A COMPUTATIONAL INVESTIGATION OF SOLUBILITY, FUNCTIONALITY AND THE ADAPTATION IN SUBCELLULAR COMPARTMENTS OF PROTEINS

A thesis submitted to the University of Manchester for the degree of PhD in the Faculty of Life Sciences

2011

Pedro Chan

Faculty of Life Sciences
# List of Contents

List of Figures ................................................................. 8

List of Tables ................................................................. 11

List of Equations ............................................................ 12

List of Abbreviations ....................................................... 13

Abstract ............................................................................ 14

Declaration .......................................................................... 15

Copyright Statement ........................................................ 16

Acknowledgements ........................................................... 17

About the Author .............................................................. 18

Rationale for Thesis Submission in an Alternative Format ....... 19

CHAPTER 1 - Introduction ................................................... 21

1.1 Proteins ........................................................................ 23
  1.1.1 Overall Description ................................................. 23
  1.1.2 Structure ................................................................. 24
  1.1.3 Protein Folding ....................................................... 26
  1.1.4 Function ................................................................. 27
  1.1.5 Stability ................................................................. 28
  1.1.6 Solubility .............................................................. 29
  1.1.7 Adaptation in Subcellular Location ......................... 31

1.2 The Cell .................................................................... 32
CHAPTER 3 - Prediction of Protein Solubility Based on Surface Charge Patches

Abstract

3.1 Background

3.2 Methods

3.2.1 Datasets

3.2.2 Calculations

3.3 Results and Discussion

3.4 Conclusions

3.5 References

CHAPTER 4 - Refining the Discrimination of Enzyme and Active Site Identification through Novel Charge-based Calculations

Abstract
4.1 Background .................................................................113

4.2 Methods .....................................................................116

4.2.1 Datasets ...............................................................116

4.2.1.1 CB Enzymes and CB Non-enzymes ......................116

4.2.1.2 Periplasmic Binding Proteins ...............................118

4.2.1.3 Enzymes ..........................................................118

4.2.2 Calculations ...........................................................118

4.2.2.1 Charge Weighting Schemes ...............................118

4.2.2.2 Electrostatic Calculations ................................119

4.2.2.3 Enzyme AS Prediction ....................................120

4.3 Results and Discussion ..............................................122

4.3.1 Enzyme/Non-enzyme Discrimination .......................122

4.3.1.1 UNI Compares with Previous Work ....................122

4.3.1.2 ISS and PKB ..................................................124

4.3.1.3 Combined Peak Potentials to Discriminate CB enzyme/CB Non-enzyme ..................................................125

4.3.1.4 Combined Peak Potentials to Discriminate ENZ/NON-ENZ sets .................................................................126

4.3.1.5 PBPs ................................................................126

4.3.2 Enzyme AS Prediction ............................................128

4.3.2.1 Compare STD to UNI and ISS ............................129

4.3.2.2 Failures for AS Location .................................131

4.4 Conclusions ...............................................................134

4.4.1 Distinguishing CB enzymes from CB Non-enzymes and PBPs ..........................................................134

4.4.2 Predicting Enzymes using Detailed Charge Distribution .....134
List of Figures

Figure 1.1: Drawing of an animal cell that represent its internal organisation ..........32

Figure 1.2: Protein-Solvent System .................................................................38

Figure 1.3: Schematic view of Coulomb’s Law ...............................................38

Figure 1.4: Schematic view the Finite Difference grid mapped over the protein-solvent system .................................................................40

Figure 1.5: Model for calculating the electrostatic energy contribution from model compound to protein ........................................................................42

Figure 2.1: Construction of the protein structural dataset ...................................58

Figure 2.2: Major ionisation zones in the pH-dependence of protein stability.........63

Figure 2.3: Normalised distribution of pH[ΔG_{FU}(min)] for protein chains in the 8 subcellular locations .................................................................................64

Figure 2.4: Subcellular pH and pH-dependent properties ....................................65

Figure 2.5: Histidine and subcellular pH..............................................................67

Figure 2.6: Histidine location and subcellular pH...............................................68

Figure 3.1: pl distribution .....................................................................................81

Figure 3.2: Schematic diagram to demonstrate calculation of ΔpKa[HIS] ...........90

Figure 3.3: Raw data for ΔpKa[HIS] in each dataset (I1, I2, B1, B2, N1, N2) ..........94

Figure 3.4: Distribution for ΔpKa[HIS] of N1 and N2 datasets. ............................94

Figure 3.5: “SumΔpKa[HIS] (by HIS)” distribution for N1 and N2 datasets. ..........95

Figure 3.6: Schematic diagram showing how histidine sidechain charge interactions may adapt to subcellular pH .................................................................97

Figure 3.7: Scatter plot with the square of correlation coefficients, comparing 2D against 1D largest non-charged patches, for N1 and N2 datasets ...............................99
Figure 3.8: Distributions of N1 (soluble) and N2 (insoluble) datasets, with respect to the largest 2D non-charged patch size. .................................................................101

Figure 3.9: ROC plot for the separation of soluble and insoluble proteins from the N1 and N2 datasets. .................................................................101

Figure 3.10: Soluble and insoluble proteins from the N1 and N2 datasets are plotted according to overall SCOP domain structural class. .................................102

Figure 4.1: A schematic diagram of the data selection for CB enzyme and CB non-enzyme sets .................................................................117

Figure 4.2: Schematic view of the charge weighting schemes ....................................119

Figure 4.3: 2D representation of the patch method (not to scale) .............................121

Figure 4.4: Comparing ENZ/NON-ENZ and CB enzyme/CB non-enzyme results in UNI .................................................................123

Figure 4.5: CB enzyme/CB non-enzyme discrimination in UNI .............................123

Figure 4.6: ROC plot for CB enzyme/CB non-enzyme discrimination in three weighting schemes .................................................................124

Figure 4.7: Cumulative plot for the discrimination between CB enzymes, CB non-enzymes and PBPs in UNI .................................................................127

Figure 4.8: The DIST of the enzyme set between different weighting schemes in XY scatter plots .................................................................130

Figure 4.9: Examples of enzymes with their AS that fall outside the 15Å threshold ...........................132

Figure 4.10: All schemes failed to predict the AS within 15Å of the centroid of 1ab8 ...........................133

Figure 5.1: Output of CRC .................................................................144

Figure 5.2: A screen shot showing the input interface .............................................145

Figure 5.3: PDB file formatting .................................................................146

Figure 5.4: Jmol displays residue conservation for 1a0iA using various E-values .........147

Figure 5.5: Display of residue conservation for 1a0iA after running CRC using an E-value of 10e-20 .................................................................148
Figure 5.6: Sequence homologues retrieved using 2ztaA via PSI-BLAST with one iteration.

Figure 5.7: Display of leucine zipper, 2ztaA.
List of Tables

Table 2.1: Subcellular locations and their experimental pHs from various literatures...61
Table 2.2: Correlations between calculated properties and subcellular pH ...............65
Table 3.1: Soluble and Insoluble protein dataset selection and filtering......................88
Table 3.2: Mann-Whitney test results for soluble and insoluble protein datasets, using calculated ΔpKaHIS values ...........................................................................................................93
Table 3.3: Mann-Whitney test P-values of features calculated for comparison of the N1 and N2 datasets .....................................................................................................................98
Table 4.1: Number of points presented in each ‘patch probe’ (10, 12 and 14 Å) used for collecting patch potentials.................................................................121
Table 4.2: Result summary for the three weighting schemes for 54 CB enzymes and 37 CB non-enzymes. ....................................................................................................................124
Table 4.3: Result summary for C-POTLs for 54 CB enzymes and 37 CB non-enzymes.................................................................................................................................126
Table 4.4: Result summary for C-POTLs for RGJW enzymes and non-enzymes set. 126
Table 4.5: Result summary for peak potentials for 54 CB enzymes and 36 PBP non-enzymes.................................................................................................................................127
Table 4.6: Parameters and results for predicting AS from the 155 enzyme set using grid step of 1.5 and 2.0 Å.................................................................128
Table 4.7: Parameters and results for predicting AS from the 155 enzyme set using grid step of 1.0 and 2.0 Å.................................................................128
Table 4.8: Enzymes with their predicted AS fall outside the 15 Å threshold. ............131
List of Equations

Equation 1.1: Coulomb’s law for calculating the electrostatic energy..........................38
Equation 1.2: Coulomb’s law for calculating the electrostatic potential.........................38
Equation 1.3: Linear Poisson-Boltzmann Equation ..........................................................39
Equation 1.4: Determination of intrinsic pKa .....................................................................41
Equation 1.5: Determination of intrinsic $\Delta \Delta G_{ienv}$ .......................................................41
Equation 1.6: Determination of $\Delta G_{FU}$ .......................................................................43
Equation 1.7: pH-dependent free energy for each nth state in the folded protein..............43
Equation 1.8: Total energy of the ionisable groups in the folded protein .........................44
Equation 1.9: pH-dependent free energy for each nth state in the unfolded protein .........44
Equation 1.10: Total energy of the ionisable groups in the unfolded protein .................44
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviations</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>amino acid</td>
</tr>
<tr>
<td>AUC</td>
<td>Area under curve</td>
</tr>
<tr>
<td>CB</td>
<td>carbohydrate binding</td>
</tr>
<tr>
<td>CRC</td>
<td>conserved residue colouring</td>
</tr>
<tr>
<td>DH</td>
<td>Debye-Hückel</td>
</tr>
<tr>
<td>EC</td>
<td>Enzyme Commission</td>
</tr>
<tr>
<td>EFLD</td>
<td>peak electric field (highest electric field on the protein surface)</td>
</tr>
<tr>
<td>ENZ</td>
<td>enzymes</td>
</tr>
<tr>
<td>FD</td>
<td>Finite Difference</td>
</tr>
<tr>
<td>FDPB</td>
<td>Finite Difference Poisson-Boltzmann</td>
</tr>
<tr>
<td>FPR</td>
<td>False positive rate</td>
</tr>
<tr>
<td>IB</td>
<td>inclusion bodies</td>
</tr>
<tr>
<td>ISS</td>
<td>ionisable same sign weighting scheme</td>
</tr>
<tr>
<td>Lg1DnonQsequence</td>
<td>Longest non-charged amino acid sequence</td>
</tr>
<tr>
<td>Lg2DnonQpatch</td>
<td>Largest non-charged surface patch</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>MSA</td>
<td>multiple sequence alignment</td>
</tr>
<tr>
<td>NON-ENZ</td>
<td>non-enzymes</td>
</tr>
<tr>
<td>PB</td>
<td>Poisson-Boltzmann</td>
</tr>
<tr>
<td>PBE</td>
<td>Poisson-Boltzmann Equation</td>
</tr>
<tr>
<td>PBP</td>
<td>periplasmic binding proteins</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>pH[ΔG&lt;sub&gt;FU&lt;/sub&gt;(min)]</td>
<td>The pH at minimal ΔG&lt;sub&gt;FU&lt;/sub&gt;</td>
</tr>
<tr>
<td>pH&lt;sub&gt;i&lt;/sub&gt;</td>
<td>intracellular pH</td>
</tr>
<tr>
<td>PKB</td>
<td>pKa buried weighting scheme</td>
</tr>
<tr>
<td>POTL</td>
<td>peak electric potential (highest electric potential on the protein surface)</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>SCOP</td>
<td>Structural Classification of Proteins</td>
</tr>
<tr>
<td>SP</td>
<td>soluble proteins</td>
</tr>
<tr>
<td>STD</td>
<td>Standard charge weighting scheme</td>
</tr>
<tr>
<td>TPR</td>
<td>True positive rate</td>
</tr>
<tr>
<td>UNI</td>
<td>Uniform charge weighting scheme</td>
</tr>
<tr>
<td>ΔG&lt;sub&gt;FU&lt;/sub&gt;</td>
<td>pH dependent folding energy between folded (F) and unfolded (U) states of a protein</td>
</tr>
<tr>
<td>ΔpKa[His]</td>
<td>histidine's pKa difference between the folded state of a protein and its model compound</td>
</tr>
<tr>
<td>ΔQ&lt;sub&gt;FU&lt;/sub&gt;</td>
<td>charge difference between folded (F) and unfolded (U) states of a protein</td>
</tr>
<tr>
<td>Δψ</td>
<td>reverse membrane potential</td>
</tr>
</tbody>
</table>
Abstract

A cell is considered to be the smallest unit of life. It carries out a variety of biochemical reactions through the activities of proteins and protein enzymes. In order to perform functions, proteins must be in their native folded state together with the correct environmental conditions. A slight change in pH or temperature could cause disruption to the electrostatic interactions within the protein, thus leading to conformational change and the loss of activity. Studies have shown that solubility could be enhanced by increasing the number of charges on the protein surface. And from the studies of extremophiles, we learned that the presence of non-polar aromatic residues could be a key for thermostable proteins. Thus, charges are important to determine the function and adaptation of proteins.

Over the decades, large amount of protein sequence and structure information relating to molecular biology has been produced. By employing algorithms, computational and statistical techniques, it is possible to analyse these data to solve biological problems. Often these investigations are based mainly on sequences since their numbers outstrip the number of available structures. However, adding structures would allow us to investigate problems such as the relationship between charges, sequence, structure and functions, which is the aim of this study.

In this thesis, the relationships between proteins and function were examined by various electrostatic features derived from charges and also geometric properties from structures. One interesting finding is that the averaged value of pH of maximum stability of proteins within a subcellular location was highly correlated to the pH of that subcellular compartment, which was due to pKas (of histidines), and their locations on the proteins. We also found that the size of the largest non-charged patch on the protein surface correlates with solubility and provides a predictor with a maximum accuracy of 76%. The use of novel charge-based methods shows little improvement in distinguishing between enzymes and non-enzymes. However, the method of using real charges with grid size of 1 angstrom has paved a way into the idea of using charges and dipoles pattern from enzyme active site to distinguish different enzymes. Finally, a web-tool for displaying conserved residues on 3D protein structure is made available to the public for identifying residues that may be of functional importance.
Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the “Copyright”) and s/he has given The University of Manchester the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

ii. Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the John Rylands University Library of Manchester. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

iii. The ownership of any patents, designs, trade marks and any and all other intellectual property rights except for the Copyright (the “Intellectual Property Rights”) and any reproductions of copyright works, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and exploitation of this thesis, the Copyright and any Intellectual Property Rights and/or Reproductions described in it may take place is available from the Dean of the Faculty of Life Sciences.
Acknowledgements

A million thanks to my parents for their love and financial support throughout my study. Many thanks to a number of pastors from the Jesuit, the Salesians of Don Bosco and the Franciscan who have been my teachers and mentors before my journey to England.

Super big thanks to Dr Jim Warwicker who has been a great supervisor. It would be impossible to walk through the scientific maze of questions and traps without his help and guidance. Special thanks to Professor Simon Hubbard for being my advisor and for his comments and recommendations. Special thanks also go to Dr Josip Lorvic for being my previous advisor. I would like to thank the BBSRC for supporting part of my research.

Lots of thanks to the group members: Salim Bougouffa, Tracey Bray, Andrew Crawley, James Kitchen and Richard Greaves for providing a fun and friendly research environment and for general discussion. Not to forget the encouragement and support from my dearest friends and relatives.

At last, but not least may God be with the ones who always have me in their prayers.

Manchester, United Kingdom
About the Author

Born in Macau. Educated in Hong Kong. Arrived the UK in 1993. Pedro has obtained his BSc in Biochemistry at UMIST in 2002 and has received his MPhil in Bioinformatics at the University of Manchester in 2005. He has been working as a lab assistant in the areas of Proteomics and Pharmacology before entering his PhD. Pedro is also interested in a range of hobbies such as acupressure, computer automation and music making. Future plan involves seeking a position that will make use of his knowledge, skills and creativity.
Rationale for Thesis Submission in an Alternative Format

Submission in the alternative format is suitable for my PhD because all the work being done can be divided into investigations with a separate emphasis. All the work presented here is original. Chapter 2 and one co-authored work (related to Chapter 5) have been published. Three pieces of work (Chapter 3 to 5) are in preparation. The alternative format will allow me to concentrate on the preparation and submission of manuscripts alongside writing the thesis. The thesis begins with a general introduction, followed by the compartmentalised work of chapters 2 to 5. A summary discussion at the end, connecting up the various parts of the thesis.
Over the past few decades rapid developments in genomic and molecular research have produced large amount of information related to molecular biology such as nucleotide sequences, e.g., European Nucleotide Archive [1]; gene expression, e.g., ArrayExpress [2]; protein sequences, e.g., UniProt Knowledgebase [3]; and structures, e.g., Protein Data Bank (PDB) [4]. This sudden increase in data is a major achievement, leading to the question of functional interpretation on a large-scale. For example, the number of protein sequences deposited in UniProtKB/SwissProt [5] was ~80k between 2000 and 2005 while ~350k were deposited between 2005 and 2010 [6]. This is in fact a challenge for biologists but then it opens up a door for scientists with computational skills. All these information need to be stored, organised, analysed and understood as well as comparing results derived from different sources. This is where Bioinformatics comes in, as a scientific discipline.

Bioinformatics is a multidisciplinary subject that employs algorithms, computational and statistical techniques and theory to analyse biological data in order to aid solving biological problems. Often, they are applied in predicting the composition, function and structure of biomolecules such as DNA, RNA and proteins. They are also involved in the development of algorithms for pattern recognition, machine learning, data mining and applied in areas such as sequence alignment, drug design and evolutionary relationship modelling. Bioinformatics has offered a quantitative way to analyse experimental data, allowing us to ask scientific questions at a molecular level and to investigate them from a large perspective. Taking the prediction of sugar binding site as an example [7], the authors perform analysis of six characteristic properties such as hydrophobicity and solvent accessible surface on the surface patches of a set of 19 annotated but non-homologous protein structures. The magnitudes of these properties were than compared amongst these surface patches, where three parameters were found to distinguish the sugar binding sites with an overall prediction accuracy of 65%. The task for calculating these atomic features and the idea of making prediction of carbohydrate binding proteins, that would have seemed impossible in the pre-genomic era, can now be accomplished in weeks, if the required datasets and programs are available. This type of analysis is only the beginning, since the post-genomic era offers many opportunities for computational and
comparative studies to contribute to biology. Thus, the investigations described in this thesis are to share some studies that contribute to the wider biological picture. The question being asked is: "How does protein structure adapt stability and function to environment and to a particular task."

In order to survive, organisms must adapt to their environments. These environments can vary widely. For example, hot springs are inhospitable to humans, while they are homes for thermophiles, organisms that can grow between 45 and 80 °C [8] and also for extreme thermophiles (hyperthermophiles), that can grow between 80 and 105 °C. For their survival, they have membranes and enzymes that can fully operate at high temperature. The underlining factors are of particular interest because they raise possibility of engineering enzymes with enhanced high temperature stability and catalytic efficiency for industrial applications [9]. Studies have suggested that the stability of thermophile-derived proteins is due to many factors. For example, (1) hydrogen bonding, as is the extent of hydrophobic interaction within protein; (2) secondary structure properties such as the amount of proline in alpha helices [10] and the beta strand content [11]; and (3) the number of charged residues in the AA composition [13, 14]. It is generally assumed that higher numbers of ionisable groups implies larger contributions to folded state stability but one study found that although on average proteins from thermophiles (and hyperthermophiles) gave a higher proportion of ionisable groups and greater folded state stability, these features do not correlate [15]. However, ionisable groups are implicated in other functions. A study of aggregation related to protein misfolding [16] found that invoke charged residues can act as "gatekeepers" for exposed beta-strands to avoid seeding amyloidosis [17]. Furthermore, an increased of the aromatic AA, Trp was observed in the non-polar surface area of hyperthermophile proteins [15]. Since carbohydrate binding sites are known to have a preference for aromatic residues (Trp in particular) [7] and sugars were found to ease osmotic stress in many extremophiles [18], it was hypothesised that the binding of such solutes could contribute to stabilising protein structure in some organisms.

These observations from thermophiles show us how proteins can adapt to challenging environments. Using the same logic, similar questions can be asked about protein adaptation within a cell. The following sections give background information relevant to how this question has been addressed in the thesis.
1.1 Proteins

1.1.1 Overall Description

Proteins are essential molecules that are vital to living cells and organisms. Thousands of chemical reactions happen in the cell which require many kinds of proteins for maintenance and growth. These proteins can be grouped into three main categories according to their roles in the cell. Protein enzymes, which belong to the soluble globular protein category catalyse a variety of biochemical reactions, other globular proteins are involved in important biological pathways such as cell signalling and cell cycle. Insoluble fibrous proteins such as cytoskeleton provide the scaffolding of a cell. Membrane proteins are responsible for cellular communications and the exchange of ions between internal and external environments. The cell (including its organelles and genome), proteins and water medium, together with the flow in and out of nutrients, ions and other metabolites form this complex system.

Apart from the involved components, various factors such as temperature, pH and ionic strength can affect the regulation of the system if one of these levels is largely changed. This may inhibit certain pathways, which in most cases result from the inactivation of proteins. Amongst these factors, solvent pH plays an important part since proteins contain several acidic and basic groups as proton donors and acceptors and their protonation states are important for folding, stability and activity of the proteins. For example, the stomach enzyme pepsin has its maximum activity at pH 2. An intake of antacids (usually a bicarbonate for treating gastric ulcer) neutralises the gastric juice, followed by an increase of pH to 6 or 7, causing deprotonation of certain acidic groups of pepsin, which lead to their conformational change and thus inactivation [19]. This sudden rise of stomach pH will eventually be restored by the secretion of more gastric juice and reactivate pepsin. This is a kind of homeostasis, as the term suggests its ability for “staying” (stasis) “the same” (homeo). Homeostasis mechanism is also found in regulating pH in cells and their organelles. Metabolites containing imidazoles of histidines are found to have pH buffering power in maintaining pH in muscle cells [95]. Since the pKₐ values of these imidazoles are close to intracellular pH, one of the two nitrogens can be protonated within the physiological pH range. Thus, the shifts of pKₐ values of histidines are crucial in pH maintenance.
In addition, the pK$_{a}$ values of ionisable groups are important in their contribution to protein stability and solubility [53]. For example, the enzyme, RNase T1 is 3.8 kcal/mol less stable when the buried Asp-76 (pK$_{a}$ = 0.6) is replaced with alanine in a mutation study [20]. The energy contribution may seem small but it is known that the energy favouring the folding of globular proteins from their unfolded forms are just 2-10 kcal/mol [21]. A protein can be less soluble at its isoelectric point, where the net charge on its surface is zero. This net charge depends on the number of ionisable groups and their pK$_{a}$ values. Therefore, introducing a charge at the protein surface trigger its interaction with water solvent.

### 1.1.2 Structure

Most of the essential structure and function of cells are mediated by proteins [22]. And from the fundamental principle of protein science, we know that protein structure is the basis for function. Since protein perform so many kinds of functions, it was once thought their structures are diverse and without regularities [23]. But under decades of research, regularities were observed and these observations were organised into a four-tiered hierarchy: primary, secondary, tertiary and quaternary structure.

A protein is a polypeptide that has undergone secondary and tertiary structure formation caused by its own interaction and with the aid of chaperones as well as post-translational modifications. The primary structure of a protein is a string of joined up AA by peptide bonds. There are typically 20 kinds of AA, each with a different sidechain. These sidechains provide distinguished physico-chemical properties allow them to play structural or functional important role in the protein. Sidechains of Ala, Val, Leu, Ile, Phe, Trp and Met are non-polar and hydrophobic, they do not form hydrogen bonds or ionic bonds with other AAs. In general, this water fearing property is the major force behind protein folding, where these non-polar groups are buried within the protein and allowed charged and polar sidechains to be exposed to solvent. Certain non-polar groups (Phe, Trp and Tyr) also have a bulky aromatic ring, which is known play an unique role in integral membrane proteins, where they are enriched at the membrane-water interface [97]. Asp and Glu are negatively charged (-1) in physiological pH, while His, Arg and Lys are positively charged (+1). Their charged nature allowed them to interact with water, thus making the protein soluble. They are also found in enzymatic site, for example, the nitrogens of
the imidazole ring of histidine play a proton shuttling part in the catalytic triad. Some AA
sidechains are also responsible for scaffolding purposes such as (1) the thiol group (-SH) of a
cysteine sidechain could form a disulphide bond with another cysteine thiol; and (2) the cyclic
structure of PRO provides rigidity to the protein structure. Secondary structures (alpha helices
and beta sheets) are local organisations of the primary sequence, which occur most frequently,
while less favorable structures such as 3_{10} helix, pi helix were also observed. Their
conformations are stabilised by hydrogen-bonding interactions. A hydrogen atom is simply a
proton surrounded by an electron cloud. When it is chemically bonded to nitrogen (N) or oxygen
(O), its electron cloud moves toward the electro-withdrawing N or O atom. Thus the positive
charge of the proton becomes unshielded. When it comes close to another atom with extra
negative charge such as an N or O, they will attract each other via their partial positive and
negative charges. Loops found at the protein surface are irregular structures, which could
possess structural or functional significance.

The level of the tertiary structure is a combination and arrangement of those secondary
structure elements. They fold to achieve the globular shape, usually with hydrophobic
sidechains packed into the protein core while charged sidechains are distributed on the protein
surface. Charged sidechains are also found in burial for participation in salt bridges. Residues
with polar sidechains often neutralise their polarity by participate in hydrogen bondings with
other polar residues or polypeptide backbone or water molecules. Within the overall protein fold,
distinct domains and motifs can be recognised [22]. Domains are usually structurally (and also
functionally) independent units, for example: pyruvate kinase (1pkn from PDB) is formed by 3
domains. By contrast, motifs are often with a functional significance, for example: the helix-turn-
helix motif, often seen in transcription factor; and the catalytic triad made up of serine, histidine
and aspartic acid. The complexity and capabilities of proteins can be extended via chemical
modification. For example, the addition of covalently attached carbohydrate can be a signal for
intracellular protein localisation; lipid modifications can help anchor a protein to the cell
membrane; and the association of metal ions or cofactors assist many enzymes in chemical
catalysis [24]. Although proteins do function as a single chain, many of them operate as a
multimer, where two or more tertiary polypeptides associate non-covalently to form a quaternary
structure. Hemoglobin is a classic example.
These basic principles about the nature of protein structure have made protein structural studies more tractable. Thus, with the available sequence and structure information, bioinformaticians are able to observe similarities or differences when making comparisons between protein families or protein sets and to report the components/features that are linked to a certain function. The following sections give brief descriptions of current knowledge that are related to protein folding, function, solubility, stability and their adaptation in subcellular locations.

1.1.3 Protein Folding

It is generally accepted that there are four types of interactions involved in protein folding and the stabilisation of the structure, which include electrostatic forces, van der Waals forces, hydrogen bonds and hydrophobic interactions. Electrostatic interactions are forces between charges, which can be described as the energy change between the distance of 2 point charges. An attractive force is usually found between water and protein surface charges for its solubility. Van der Waals forces are weak interactions occur between all atoms, either polar or non-polar [98]. They include forces between dipole (permanent and/or induced) molecules, whereas hydrogen bonds are abnormally strong dipole-dipole attractions that involved molecules with -OH, -NH in proteins. Their main contribution to protein folding is to stabilise the secondary structures. The folding of protein into its native conformation was first believed (1) to be a spontaneous reaction guided by hydrophobic interaction and (2) the protein shape is encoded in the AA sequence, based on experimental observations for ribonuclease in 1960s. They were later proved to be partly true, since the folding for certain proteins with longer chains (> 100 AAs) are found to be assisted by chaperones. For small proteins (< 100 AAs), two levels of foldings were observed: unfolded and folded state. Three steps can be observed in larger proteins, with an intermediate state called molten globule via a process called hydrophobic collapse (hydrophobic sidechains suddenly clump together), or with fully formed secondary structures as the intermediate state.
1.1.4 Function

Function is everything that happens to or through a protein [25]. In order to perform functions, proteins usually interact with other molecules and these interactions include such as the binding of ligands in receptor sites, allosteric binding, the binding of antibodies to antigens, protein-DNA interactions, protein-protein interactions and multimerisation [26]. The shape and the chemical properties on the protein surface are the key factors for these interactions. Early studies have shown clefts are relevant to binding sites [27]. It has been suggested that the active site usually lies in the largest cleft of the protein, and smaller sites may contribute in the binding of allosteric effectors [28, 29]. In a study of 67 enzyme-ligand structures, which computes their cleft volumes, the authors found that the ligand is bound in the largest cleft in over 83% of proteins [26], which provides a simple way to locate binding sites.

Being enzymatic is considered as a function of a protein and the identification of the largest cleft is amongst one of the features that makes up an enzyme. Having a strained environment can also be an indication of being an enzyme. Study of the mutation of catalytic residues in T4 lysozyme gave reduced enzymatic activity but increased stability [30]. This highlights that the ground state of the enzyme is strained, which is caused by an environment with large shift of pKa that is often associated with catalytic residues [31, 32]. Estimation of these strain and perturbed energetics can be obtained by electrostatics calculations. Other characteristics are found in the proteases. A study of 36 evolutionarily unrelated proteases found that their surface areas are smaller than average and they possess higher Cα densities (more tightly packed) [33]. They also found that proteases have fewer helices but more loops when compared to non-proteases. In a later study, the authors have found that on average O-glycosidases have larger negative surface patches than non-O-glycosidases [34]. Higher frequency of aromatic residues: Trp, Tyr and Asn are also found in O-glycosidases, these residues are known to be common docking sites for cyclic carbohydrate molecules [35].

It has been shown that simple attributes such as secondary structure content, AA propensities, surface properties and ligands were able to identify top EC (Enzyme Classification) class for an enzyme [36]. Recent work from the same group [37] has shown features that are observed in individual EC class. Oxidoreductases have more non-polar active sites attributed to cofactor binding. Lyases form a high proportion of oligomers than any other EC classes, while
hydrolases form the largest proportion of monomers. In addition, large amounts of protein signatures that link to protein function can be found in InterPro [38], which is an integrated database for sequence patterns from databases with different focuses such as PRINTS [39] (focus on protein families), Pfam [40] (focus on divergent domains), PROSITE [41] (focus on functional site). These motifs are usually short and conserved sequences that are of high utility in the problem of function prediction.

1.1.5 Stability

The function of a protein also depends on its stability. In terms of thermodynamics, protein stability depends on the difference between the free energy of the folded and unfolded state. After a protein is synthesised in an aqueous environment of the cell, it is guided by the thermodynamic forces to fold into its folded native state with its free energy just marginally lower than its unfolded state, usually in a range of 2 to 10 kcal/mol for globular proteins. Thus, protein stability could be enhanced either by increasing the stability of folded protein, or by decreasing the stability of the unfolded form, or by a combination of both [42]. Experimental studies using mutagenesis to study the effect of residue substitution in proteins allow us to gain insights of the interactions being involve in protein stability. It is generally agreed that the hydrophobic effect is the major factor in stabilising folded globular proteins. The effect of hydrophobicity was demonstrated in a study of four different proteins with a total of 54 single AA substitutions [43]. Substitution of the non-polar residue within the protein core was replaced by a smaller non-polar residue, for example: Ile is substituted by Val and Leu is substituted by Ala. The resultant change between the free energy of the folded and unfolded state of the protein (∆∆G) is taking as a measure of the difference between hydrophobic stabilisation provided by the two AA [43]. Amongst these results, 17 different leucine-to-alanine substitutions have shown energy changes that range from 1.7 to 6.2 kcal/mol, which indicates the importance of hydrophobicity in protein stability.

Electrostatics also play a role in protein stability. The non-covalent interactions induced by charged and polar residues are known to be fairly weak interactions, but these small stabilizing interactions can add up to make an important contribution to the overall stability of a protein [44]. Early experimental results showed long-range electrostatic contributions are relatively
small in a mutagenesis study of T4 lysozyme, which has 9 positive charges at neutral pH [45]. Five positively charged surface residues (Lys or Arg) were individually replaced with glutamic acid together with eight selected double, triple, and quadruple mutants as to reduce sequentially the out-of-balance formal charge on the protein from +9 to +1. The thermal stabilities of all variants were fairly close to the wild-type ranging from 0.5 kcal/mol more stable to 1.7 kcal/mol less stable. For salt bridges, it has been speculated that the burial of charges is a means to reduce protein stability [46]. One hypothesis is that the stability of the hydrophobic core may be modulated by electrostatic interactions and repulsion [47]. Salt bridges may also increase the specificity of folding because there are fewer ways to organise the charge groups while there are more ways for the packing of hydrophobic residues [47, 48]. Moreover, a suitable electrostatic environment is required for enzyme catalysis and/or protein-ligand binding, therefore the organisation of charged groups are important.

1.1.6 Solubility

The majority of proteins, especially enzymes function in the aqueous environment of the cell (except proteins that reside within the plasma membrane such as transmembrane proteins and integral membrane proteins), their degree of solubility is of primary importance, usually a high solubility is required for their functionality. Protein solubility is a function of many factors, such as pH and temperature. Generally, proteins are more soluble in acidic or basic pH because excess charges of the same sign, causing charge repulsion among molecules and water, contribute more to solubility. On the other hand, proteins are least soluble at their isoelectric point (pI). At this pH, the overall charges of a protein is zero, the electrostatic forces are cancelled out between opposite charges, which make less interaction with water, thus favour proteins to approach each other and aggregate. Temperature also influences solubility. Proteins can be denatured by high temperature. The non-covalent interactions/bonds such as hydrophobic interactions, hydrogen bonds, salt bridges are destabilised, causing some secondary and tertiary structures to unfold. Thus, the interactions with water are reduced by certain hydrophobic groups, for example the thiol (-SH) in cysteine [49].
Without a general theory of protein solubility, it is difficult to determine how a specific distribution of charges affects solubility [50]. Part of our understanding about solubility mostly came from the experimental studies of RNase Sa. RNase Sa is a small acidic protein (96 AA) with a pI of 3.5. One study compared the solubility between its native and denatured state [51]. The authors found that the solubility increases sharply from its pI towards both pH ends (within pH 3 and pH 4.5) in the native state, while in denatured state solubility rise slowly from its pI towards both pH ends (within pH 2 and pH 6.5). This indicates structure is crucial to maintain solubility. In a later study, the authors measured the solubility of RNase Sa with Thr-76 (the most exposed sidechain) being replaced by the other 19 AA one at a time. They found that His, Asn and Gln make unfavourable contribution to solubility while Asp and Ser have the opposite effect [52].

Because many proteins are now used as drugs, solubility is a concern in the biopharmaceutical industry as well as in the area of biochemical studies to characterise new proteins [53]. Poor solubility during over-expression of recombinant proteins in E.coli is a common problem, this includes formation of inclusion bodies and low protein yield. Over the decades, several groups have developed methods to predict protein solubility, initially built as a high-throughput screening system for selecting structural genomics (SG) proteins that do not form inclusion bodies. SG proteins are proteins from the whole proteome of an organism that are of medical importance to human such as M. tuberculosis (TB), which causes tuberculosis. The structures of these proteins are being solved by the TB Structural Genomics Consortium. An early prediction model, which analysed six features including average charge, turn-forming residue fraction, cysteine fraction, proline fraction, hydrophobicity and total number of residues on a set of 81 proteins with an accuracy of 88% found that average charge (Asp, Glu, Lys and Arg) and turn-forming residue (Asn, Gly, Pro and Ser) fraction have a strong correlation to solubility in proteins from overexpression [54]. A slightly revised version of this model [55] was widely considered to be the most accurate method. There has been some discussion of the accuracy of these models, and further studies have been undertaken [56-58]. Some of these studies have been limited by questions over dataset size and redundancy.
1.1.7 Adaptation in Subcellular Location

As already mentioned proteins are very diverse, most of them function inside the cell and in different subcellular compartments, while some are exported outside of the cell and reside in body fluids. The condition of these compartments vary, however there are common features in all cell type. For example: (1) when compared to exterior environment, the interior of the cell has a higher overall concentration of ions, small metabolites and proteins [59]; (2) All cells actively export anions to counteract osmotic expansion, thus making the intracellular pH slightly alkaline; (3) A voltage difference is also maintained across the plasma membrane, creating a higher concentration for certain ions (H+, Na+) in the extracellular environment [59]. Values of pH in other subcellular compartments are also tightly regulated. pH is nearly neutral in ER and mitochondria. Vacuoles, lysosomes and Golgi are more acidic while nucleus and peroxisomes are more basic. For adaptation, the charged residues of proteins might have co-evolved with these physiochemical environments for structural or functional purposes since proteins in the same location experience the same constraints, therefore global features are likely to be presented. An example would be the pI of proteins, which describes the balance between acidic and basic residues and has long been a standard measure for distinguishing between proteins. Previous searches for correlations between subcellular localisation and pI have failed to detect meaningful correlation [60]. However, we found that the average pH of maximal stability for proteins in a subcellular compartment correlates better with subcellular pH than pI [61]. Other global features such as the total AA composition and variants on this theme (e.g. dipeptide composition) have also been reported to distinguish proteins from different subcellular locations [59, 62-65] but there are no biological grounds for why such a correlation exists [66]. In addition, targeting peptides are well known sequence feature that are found in mitochondrial, chloroplastic and most secretory proteins as an indicator for their subcellular locations.
1.2 The Cell

1.2.1 Internal Organisation

The cell is regarded as the basic and smallest unit of life and all living things are composed of one or more cells. The size and the internal component of a cell depends on the type of organisms and the type of cell. For example: (1) the cell size for an eukaryote varies between 10 and 100 micrometres in diameter, which is 10 times bigger than the archaea or bacteria cell, and the later ones do not have nucleus, mitochondria or any membrane bound organelles; (2) red blood cells in mammals do not have a nucleus. Figure 1.1 shows a drawing of an animal cell.

![Figure 1.1: Drawing of an animal cell that represent its internal organisation](image)

A typical cell is enclosed by its plasma membrane, which is a lipid bilayer composed of phospholipids that act as a barrier to keep essential proteins, ions and small molecules in place and to avoid other molecules from entering. There are membrane proteins either penetrate the lipid bilayer or associate on the inside or outside of the plasma membrane. They function as transporter, receptors or responsible for cell adhesion. A cell wall may be presented on the exterior of the lipid bilayer if it is a plant cell. Its function is to withstand osmotic pressure from the difference in solute concentration between the cell’s internal and external environment [67].
The interior of the cell is the cytoplasm, which is a site for multiple cell processes. It is made up of mainly water (~70% of the volume) [68] with a high concentration of potassium ions, a low concentration of sodium ions, a trace amount of other ions such as chloride, magnesium and calcium [69]. About 20-30% of cytoplasm is occupied by proteins [70]. A single nucleus is situated in the cytoplasm, where DNA, the genetic material is located. Near the nucleus are the structures of rough endoplasmic reticulum (ER), the site for protein synthesis and the smooth ER with functions such as synthesis of lipids and steroids, metabolism of carbohydrate and attachment of receptors on cell membrane proteins. The Golgi apparatus is responsible for sorting, modifying and packaging of macromolecules such as proteins and lipids before processing them for secretion. The waste materials in the cell are broken down by the acidic organelles, lysosomes, which contain enzymes for the digestion of macromolecules. The same roles are performed by lytic vacuoles in yeast and plant cells [71]. Most mature plant cells are occupied by a single large vacuole, which takes up more than 30% of the cell’s volume [72]. In plants, the vacuole performs a variety of functions, for example: the isolation of harmful materials, maintaining internal hydrostatic pressure within the cell and the maintenance of an acidic internal pH. Though vacuoles are also presented in some animal cells, their roles are being subordinate, assisting the processes of exocytosis (e.g. expels cellular waste) and endocytosis (e.g. receive nutrients). The breakdown of very long fatty acids chains are the major role for peroxisomes in yeast and plants, while in animal cells the breakdown process is shared with mitochondria [72].

The mitochondrion is the power factory of the cell that generate adenosine triphosphate (ATP) as an energy source. It is a double membrane organelle with five distinct compartments, each carries out specialised function. The outer membrane act as a barrier that only allows molecules of up to 5000 Daltons to diffuse in and out. A signal peptide is required for the import of large proteins. The intermembrane space is between the outer and inner membrane, it contains concentrations of molecules such as ions and sugars similar to the cytoplasm. The matrix in the inner membrane contains proteins to perform functions such as: enzymes for oxidative phosphorylation, ATP synthase to generate ATP in the matrix and transport proteins that regulate the passage of metabolites in and out of the matrix [72]. The matrix also contains copies of the mitochondrial genome and its own tRNA and ribosomes. The inner mitochondrial membrane is compartmentalised into several cristae, which increase its surface area to
enhance ATP production. Mitochondria are also known to contribute to the homeostasis of calcium in a cell [73]. Like the mitochondrion, chloroplast is the power source for photosynthetic organisms such as plants, which generate energy by converting light in the form of ATP and reduce NADP to NADPH. It is a double membrane organelle with certain compartments. The fluid found within the chloroplast is called stroma, which contains its own circular DNA, ribosomes and some proteins. Thylakoids stacks are presented within the stroma and are the sites for photosynthesis.

1.2.2 pH Homeostasis and Proton Buffering

All living beings, from the simplest prokaryotes to complex multicellular organism are constantly battling against the environments. The maintenance of an appropriate pH within membrane-enclosed compartments is one of the universal challenge across all organisms [74]. Over the past 50 years, intensive studies have been carried out in neutralophiles, acidophiles and alkalophiles, and a variety of pH homeostatic mechanisms have been identified [75]. For example, in acidophiles, (1) a highly impermeable cell membrane can restrict the rate of proton influx into the cytoplasm; (2) the reverse membrane potential (Δψ) can partially deflect the inward flow of protons, one suggested mechanism is by a greater influx of potassium ions, which creates positive Δψ (as opposed to negative Δψ in neutralophiles) as a chemiosmotic barrier against the proton gradient [76]; (3) comparative genome analysis suggested that a larger proportion of DNA and protein repair systems might be present to help sustain life at low pH. In neutralophiles such as *E. coli*, the presence of enzymes and/or chemicals that are capable of buffering can help maintain the intracellular pH (pH\textsubscript{i}), for example: buffer molecules that have basic AA (Lys, His and Arg) can sequester protons; phosphoric acid (H\textsubscript{3}PO\textsubscript{4}) can perform three proton dissociations into a dihydrogen phosphate anion (H\textsubscript{2}PO\textsubscript{4}\textsuperscript{-}; pKa = 2.12), hydrogen phosphate anion (HPO\textsubscript{4}\textsuperscript{-}; pKa = 7.21) and orthophosphate anion (PO\textsubscript{4}\textsuperscript{3-}; pKa = 12.67). At near-neutral pH, HPO\textsubscript{4}\textsuperscript{-} can accept or lose a proton without affecting the pH of the solution. In addition, active proton pumping is found in all organisms to maintain pH gradient.

Moreover, a few studies have showed that the influx of H\textsuperscript{+} into the cell only causes minimal effect to the pH\textsubscript{i} in spite of the inhibition of H\textsuperscript{+} pumps [77-79]. This indicates there are buffering power existed in the cytoplasm, although not all buffering molecules are known. It was
suggested that amino acid sidechains from proteins were primarily responsible for buffering with a decrease in pH\(_i\) of \textit{E. coli} [77] and in acidophiles [80]. Evidence for the cytoplasmic buffering power in acidophiles was demonstrated by comparing the estimated pH\(_i\) (in an unbuffered environment) to experimental measured pH\(_i\) under the addition of H\(^+\) influx inducing agent to a cell suspension. The buffering power of \textit{T. acidophilus} was studied [80]. \textit{T. acidophilus} is an acidophile bacterium with an optimal growth pH of 3 and a cytoplasmic pH of 6. At its optimal growth pH, it can respire and expel H\(^+\) immediately through the respiratory pumps if there is a drop in the external pH. The respiration pumps can be inhibited by a low external pH (pH 1.5 or under), anaerobiosis (lack of oxygen), or treatment with inhibitors such as azide or cyanide. The organism cells containing 38 mg of cell protein in a total volume of 8 ml. The total cell water volume of the cell suspension is 75.8 µl. The cell water determination was described by Rottenberg (1979) [100], where 1 mg of cell protein corresponded to 2.04 ± 0.29 µl [99], therefore 2.04 µl × 38 = 77.52 µl, which shows how 75.8 µl was derived. External pH was shifted from 3.015 to 3.951 after the addition of H\(^+\) influx inducing agent and the respiration pumps are inhibited by anaerobiosis. Then, \(6.8 \times 10^{-6}\) mol of H\(^+\) was estimated to be taken up by the cell fluid, generating an internal H\(^+\) concentration of \(8.9 \times 10^{-2}\) mol of water H\(^+\) per litre, equivalent to pH 1 in the estimation. However, the measured pH\(_i\) is only reduced from pH 6.0 to pH 5.2. Internal pH was measured by the transmembrane distribution of aspirin, using the flow dialysis technique [99]. The estimated buffering capacity is 100 nmol of H\(^+\) per mg of protein per pH, and an experimental one is further checked by direct measurement using a new set of disrupted cells. The extracts were titrated from pH 6 to 5, with an average value of 97 ± 41 nmol of H\(^+\) per mg of protein per pH. For comparison, the buffering capacity of \textit{E. coli} was found to be 33 nmol of H\(^+\) per mg of protein per pH when the extracts were titrated from pH 6 to 5. These data clearly demonstrated H\(^+\) ions flow into the \textit{T. acidophilus} cells down their concentration gradient. Their buffering capacity of 100 nmol of H\(^+\) per mg of protein per pH unit, which is sufficient to account for the relative homeostasis of the pH\(_i\) under metabolically inactive conditions [80].
1.3 Protein Electrostatics

Protein electrostatics is the study of interactions and effects of stationary electric charges in and between proteins, solvent and ions. These interactions together with the charge state (0, +1 or -1) of the ionisable groups can be used to estimate the pKa values, folding free energy and electrostatic potential of a protein. Using the appropriate algorithms and parameters, predicted pKa value of an ionisable group could make reasonable agreement with experimentally determined pKa value [82-84]. This was a crucial breakthrough as it allows the study of pH-dependent properties of thousands of protein using atomic coordinates from their structures. Currently, there are more than 60000 protein structures available from the RCSB Protein Data Bank [4]. This valuable resource allows us to ask more general questions about proteins as well as precise details for individuals.

1.3.1 Development of Continuum Model

Theories of continuum models for proteins were first described by Linderstrøm-Lang, Tanford and Kirkwood [84, 85]. In those days the protein and solvent were treated as low and high dielectric material respectively. A dielectric constant of 20 or less was assigned to the protein while a dielectric constant of ~80 was assigned to the surrounding solvent. Charges of the ionisable groups were treated as if all were distributed on the protein surface. The resulted ionisable pKas were far from the experimental measured values due to the exclusion of the locations of the ionisable groups. The resultant $\Delta G$, folding free energy of the protein was repulsive (not favour in folding) at all pHs (except at the isoelectric point) because a simple estimation was used ($\Delta G$ is proportion to the power of net surface charge). The estimation of the pH-dependence was improved by representing secondary structures of protein by cylinders and slabs [86]. The major breakthrough was the introduction of finite difference (FD) technique [87] to solve the Poisson-Boltzmann (PB) equation using atomic co-ordinates from protein structure. Although FD provided a very detailed estimation for electrostatic potential on any point of the protein surface, the use of different dielectric values from different researchers could not agree with the final $\Delta G$. This is because a uniform dielectric values exclude the effect of large pKa shift for restricted ionisable groups in some burial locations. A few of these small energies differences (5-10 kcal/mol) may give rise to a large difference when compared to the
1.3.2 Coulomb’s Law

When talking about electrostatics of proteins, we usually mean those features that are important to our understanding of proteins such as the electrostatic potential, folding free energy and pKa of the sidechains of ionisable groups. Their estimation requires solving the Poisson-Boltzmann equation (PBE) for a protein-solvent system in a continuum model treating protein and water as uniformly smooth materials (Figure 1.2). Protein is assigned with a low dielectric (usually between 2 and 6) since the positions of the polar groups and side chains are generally restricted and cannot polarised in response with an external dielectric field. Water is assigned with a high dielectric (around 80) since their dipoles can orient freely. The PBE calculates the electrostatic potential of a point charge and can be perceived as a complex version of Coulomb’s law taking more interactions from the surroundings into account. Coulomb’s law and its derivative below describing the energy and potential of two point charges in a vacuum system with an infinite boundary (Figure 1.3). The electrostatic energy between the pair of points, \( i \) and \( j \) with their charges, \( q_i \) and \( q_j \), separated by a distance, \( r_{ij} \) can be calculated by equation 1.1. The electrostatic potential, \( \phi_i \) at point \( i \) by charge, \( q_i \) is given by equation 1.2.
Figure 1.2: Protein-Solvent System

\[ U_{ij} = \frac{q_i q_j}{4\pi \varepsilon_0 r_{ij}} \]  
Equation 1.1

\[ \varphi_i = \frac{U_{ij}}{q_i} = \frac{q_j}{4\pi \varepsilon_0 r_{ij}} \]  
Equation 1.2

\[ U_{ij} \] = Electrostatic energy between 2 charges: \( q_i \) and \( q_j \)

\[ \varphi_i \] = Electrostatic potential at point \( i \) by charge, \( q_j \)

\[ r_{ij} \] = Distance between the centres of \( q_i \) and \( q_j \)

\[ \varepsilon_0 \] = Dielectric constant in vacuum

Figure 1.3: Schematic view of Coulomb’s Law
In addition, the potential of any point in space from multiple point charges is the sum of each potential from the source charges to the point. Moreover, if the charges are in a medium instead of vacuum, $\varepsilon_0$ is replaced with $\varepsilon_0 \varepsilon_r$, where $\varepsilon_r$ is the dielectric of the medium. Since equation 1.2 can only model a uniform dielectric in an infinite medium, we need the more complicated PBE to solve our protein-solvent system. The linear PBE (Equation 1.3) is formed by the Poisson equation and Boltzmann’s distribution, which accounts for charge density in the system and describes the charge density for solvent respectively. It is used to model simple shape such as protein sphere. For complex protein shape with atomic details, a numerical solution such as the FD method is applied to solve the PBE or possibly the non-linear PBE.

$$\nabla \cdot [\varepsilon(r) \nabla \phi(r)] - \varepsilon(r) \kappa^2(r) \sinh[\phi(r)] + \{4\pi \rho^p(r) - \kappa^2 \phi(r)\}/kT = 0$$

Equation 1.3

- $\varepsilon(r)$ = dielectric constant
- $\phi(r)$ = electrostatic potential
- $\varepsilon(r) \kappa^2(r) \sinh[\phi(r)]$ = salt effect (includes ionic strength)
- $4\pi \rho^p(r)$ = charge density in protein
- $\kappa^2 \phi(r)$ = charge density in solvent
- $k$ = Boltzmann constant
- $T$ = absolute temperature

The protonation state of a protein is required before solving the PBE. This is because hydrogen atoms are often missing due to low resolution in their crystal structures. pH-independent hydrogen atoms can easily be assigned by simple algorithm but hydrogen atoms for titratable sites require statistical method such as Monte Carlo (MC) to determine their protonation states by weighting out the more energetically favourable configurations as a function of pH because these sites depend on all other sites. MC sampling of protonation states have been employed by a number of studies [88, 89].
1.3.3 Finite Difference

In FD, the protein-solvent system is mapped onto a Cartesian grid (x, y, z). The grid boundary is based on the solvent probe with radius of 1.4 angstroms. The electrostatic potential, charge density, and position dependent dielectric constant of each charged atom of the protein within a grid cube is replaced by the central grid point. These electrostatic terms are influenced by the six nearest grid points. The calculation can be iterated to convergence and an approximation obtained. These are summarised in figure 1.4.

![Figure 1.4: Schematic view the Finite Difference grid mapped over the protein-solvent system](image)

(a) Protein-Solvent System (Protein in metallic green; Solution in black); (b) Part of the solute-solvent placed in to the 3D grid (represents as 2D in here; a histidine sidechain is in a ball & stick representation); (c) A close up of a nitrogen atom of histidine; (d) A grid with the central grid point and its 6 nearest neighbours.
1.3.4 The pKa Calculation

Ionisable groups play crucial roles in biological processes via their ability in accepting/donating protons and/or interacting with other charged/polar parts of other macromolecules. pKa is a measure of this ability as well as a reflection of the free energy difference between the neutral and charged state of its ionisable group [90]. Therefore a precise estimation of the pKa is important but the calculation could be complicated by model approximation and multiple conformation of the sidechain. Methods developed for pKa calculation are useful when applied to regions with limited solvent accessibility. The overall pKa analysis would be unreliable when a fixed protein dielectric is based on a single sidechain conformer.

Usually, such titratable groups are sidechains of certain AA and the protein termini. Their model compound pKas are: Asp 4.0; Glu 4.4; His 6.3; Lys 10.4; Arg 12.0; Cys 8.3; Tyr 10; amino terminus 7.5; carboxyl terminus 3.8. In our case, a model compound pKa represents the proton binding affinity of the sidechain of an isolated AA and the pKa defines at what pH the titratable group should become charged or neutral. The actual pKa of a titratable group is very likely to be differ from its model compound pKa when it is part of a folded protein. Whether it is buried inside or on the protein surface, it's pKa will be affected by nearby charges of other titratable groups or polar residues (Figure 1.5). The intrinsic pKa (pKa\textsubscript{int}) of a titratable group in a protein (as the pKa it would have if all other titratable groups were neutral) can be described by:

$$pKa\textsubscript{int} = pKa\textsubscript{0} - \gamma(i) \frac{\Delta\Delta G\textsubscript{env}}{2.3kT}$$  \hspace{1cm} \text{Equation 1.4}$$

where pKa\textsubscript{0} is the model compound pKa in solution, $\Delta\Delta G\textsubscript{env}$ is the change in electrostatic energy of charging the ionisable group in the all neutral protein environment relative to the same charging process of the isolated group in water [102]. $\gamma(i) = -1$ or $+1$ for an acidic and basic group, respectively. Also, $\Delta\Delta G\textsubscript{env}$ is a combined term for:

$$\Delta\Delta G\textsubscript{env} = [\Delta G\textsubscript{Born}(\text{protein}) - \Delta G\textsubscript{Born}(\text{model}) + \Delta G\textsubscript{back}(\text{protein}) - \Delta G\textsubscript{back}(\text{model}) + \Delta G\textsubscript{interact}]$$  \hspace{1cm} \text{Equation 1.5}$$
where $\Delta G_{\text{Born}}$ is the Born solvation energy (for energy changes from solvent to protein environment). $\Delta G_{\text{back}}$ is the charge interaction between charged groups and permanent dipoles. $\Delta G_{\text{interact}}$ is the interaction between ionisable group $i$ and other ionisable groups. The Born energy differences are zero for DH calculations [92]. Since the interaction between titratable groups is pH-dependent, all titratable sites would have different ionisation state at different pH. Therefore it would be appropriate to calculate the energy for each of the possible ionisation states of the protein, and use these energies to evaluate the partition function for these states at a range of pH-values. A protein with 25 titrating residues has $2^{25} = 33554432$ possible ionisation states, it would be computationally expensive if more than 30 titrating residues were considered. This is because on average, a protein would have around 400 residues, given that around 26% are ionisable groups [93], there will be about $2^{100}$ possible ionisation states. Thus, a MC sampling method was employed to determine the ionisation states of many interacting titratable residues. The MC method used 5000 steps, weighting the more energetically favourable configurations. For energy values at pH 0, where the protonated state is heavily predominant, a ‘reduced sites’ method that restricts the number of ionisable groups considered was used [94].

**Figure 1.5: Model for calculating the electrostatic energy contribution from model compound to protein**

The dotted line represents the dielectric boundary. Figure after Bashford & Karplus, 1990 [94].
1.3.5 The free energy term (ΔGFU)

In chapter 2, we calculated pH[ΔGFU(min)] (pH where the protein has its lowest folding free energy) for each protein structure in our dataset. ΔGFU is calculated at every 0.5 pH unit between pH 0 and pH 14. pH[ΔGFU(min)] is determined at the pH where ΔGFU is at its minimum, i.e. the most negative value. ΔGFU is defined as the energy difference between the folded (F) and unfolded (U) states of protein as a function of pH, which includes the energy contribution of electrostatic energy (pH-dependent) of the ionisable groups, and contributions (pH-independent) from Born solvation energy, and the interactions between charged groups and permanent dipoles. The estimation of ΔGFU made in our study follows the description by Antosiewicz and co-workers [96] and is given by:

$$\Delta G_{FU} = \Delta G_{ion}(F) - \Delta G_{ion}(U)$$

Equation 1.6

where ΔG_{ion}(F) and ΔG_{ion}(U) represent the total free energy of the ionisable groups in the folded and unfolded states respectively. For a given protein structure, the initial step is to calculate the intrinsic pKa for each ionisable group (pKa^{int}), which have been covered (See equation 1.4 and 1.5 in sub-section 1.3.4). These intrinsic pKa values are used to determine ΔG_{ion}(U) and ΔG_{ion}(F). For a protein with N ionisable groups, there are 2^N possible ionisation states (n). The pH-dependent free energy for each n-th state in the folded protein (ΔG^n(F)) is given by:

$$\Delta G^n(F) = \sum_{i=1}^{N} \{ \delta_n(i)\gamma(i) \cdot 2.3kT(pH - pKa^{int}_i) + \sum_{1 \leq j \leq i} \delta_n(i)\delta_n(j)\Delta G^{ij} \}$$

Equation 1.7

where the vector, δn(i) is 0 when group i is neutral and 1 when it is charged. ΔG^{ij} is the electrostatic interaction between group i and j, in their charged state [101, 102]. And the total free energy of the ionisable groups of the folded protein is given by:
where $\sum \exp [-\Delta G_n^i(F)/kT]$ is the Boltzmann weighted sum of the $2^N$ ionisation states of the folded protein.

An identical set of equations can be used for determining $\Delta G^i_{\text{ion}}(U)$ but in the unfolded state, the ionisable groups have no single conformation. It is assumed that the pKas of the ionisable groups are identical to the model compound pKas, where all groups are exposed to solvent and do not interact to each other electrostatically. Thus, the pH-dependent free energy for each $i$th state in the unfolded protein ($\Delta G^i(U)$) is given by:

$$\Delta G^i(U) = \sum_{i=1}^{N} \{ \delta_n(i)\gamma(i) 2.3kT(pH-pK_{\alpha_i}^0)\}$$  \hspace{1cm} \text{Equation 1.9}$$

and the total free energy of the ionisable groups of the unfolded protein is given by:

$$\Delta G^i_{\text{ion}}(U) = -kT \ln \left( \sum_{n=1}^{2^N} \exp [-\Delta G_n^i(U)/kT] \right)$$  \hspace{1cm} \text{Equation 1.10}$$

Equation 1.4 to 1.10 describe how $\Delta G_{FU}$ is obtained but calculating for all ionisation states are computational demanding, therefore a computational algorithm for random sampling such as Monte Carlo sampling method is used for selecting the more energetically favorable configurations (See sub-section 1.3.4).

In the current study, we employed algorithms to calculate the pH-dependent of maximum stability of protein as one of the features. These molecular interactions were taken into account and were modeled in the equations. However, our model does not account entropic effects, i.e. hydrophobic effects, arising from solute-imposed constraints on the organisation of water. This is because our model rests on simple idea that non-polar of a solute tend to cluster together, whereas polar and charged groups tend to remain in water.
1.4 The Current Investigation

Studies we have described so far have provided an overview of what we know about proteins via experimental and computational studies as well as the limits of these approaches. We have developed different methods to help improve current methodology. Sequence/pattern homology based methods are generally more popular than structural based methods since the number of available sequences far outstrip the number of structures. Although much can be achieved with studies of sequences alone, adding structures allows other problems to be investigated, for example, the stability/structure/function relationships. Our initial thought is that we have plenty of tools to calculate various features with proteins. Therefore, we would like to make use of our tools to investigate whether structural derived properties could provide any clues to the function of proteins or observations related to the subcellular environments of proteins.
1.5 References


73. Koeppen AH. Basic Neurochemistry:: Molecular, Cellular, and Medical Aspects; edited by GJ Siegel, BW Agranoff, RW Albers, SK Fisher, and MD Uhler, Lipincott-Raven,


CHAPTER 2 - Evidence for the Adaptation of Protein pH-dependence to Subcellular pH

PAPER 1

Pedro Chan ¹ (PC), Jim Warwicker ¹* (JW)

¹Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, M13 9PT, UK

*Corresponding author

JW and PC together designed the study and wrote code for the calculations. The dataset was assembled by PC, and PC performed the calculations and most of the data analysis. JW and PC together wrote the paper.

Published in BMC Biology, Volume 7, Issue 9, 22 Oct 2009
Abstract

The availability of genome sequences, and inferred protein coding genes, has led to several proteome-wide studies of isoelectric points. Generally, isoelectric points are distributed following variations on a biomodal theme that originates from the predominant acid and base amino acid sidechain pKas. The relative populations of the peaks in such distributions may correlate with environment, either for a whole organism or for subcellular compartments. There is also a tendency for isoelectric points averaged over a subcellular location to not coincide with the local pH, which could be related to solubility. We now calculate the correlation of other pH-dependent properties, calculated from 3D structure, with subcellular pH.

For proteins with known structure and subcellular annotation, the predicted pH at which a protein is most stable, averaged over a location, gives a significantly better correlation with subcellular pH than does isoelectric point. This observation relates to the cumulative properties of proteins, since maximal stability for individual proteins follows the bimodal isoelectric point distribution. Histidine residue location underlies the correlation, a conclusion that is tested against a background of proteins randomised with respect to this feature, and for which the observed correlation drops substantially.

There exists a constraint on protein pH-dependence, in relation to the local pH, that is manifested in the pKa distribution of histidine sub-proteomes. This is discussed in terms of protein stability, pH homeostasis, and fluctuations in proton concentration.
2.1 Background

The post-genomic era allows many basic questions to be addressed, such as the nature of the biological components [1], control of expression levels for these components [2], their interaction networks and dynamics [3], and the ultimate realisation of metabolic function from the parts [4]. Even for proteins, the most studied nodes of biological interaction and function, there is much to discover about how form relates to function. Differences in the characteristics of amino acid sidechains, and in the stoichiometry of their incorporation into proteins, means that physico-chemical properties of proteomes and sub-proteomes can be variable. Several reports of proteome-wide properties have appeared. Features studied include amyloidogenic potential and biological context [5], propensity for disorder and protein degradation [6], amino acid composition and gene expression [7], protein targeting and N-terminal features [8], and the inclusion of physico-chemical properties into proteome browser resources [9].

Protein isoelectric point (pI) can be conveniently estimated from amino acid sequence. Three-dimensional structure gives rise to charge interactions that are important in considering protein folded state stability, but generally give small changes in pI compared with linear sequence [10]. Several groups have looked at computed proteome-wide pI distributions, with the outcomes falling into two overall categories. First, there has been discussion of the form of the pI distribution with pH [11-13], and demonstration that a general bimodality results from the predominant acidic and basic pKas of the Asp/Glu and Lys/Arg sidechains. Second, the relative populations of peaks (acidity versus basicity, or pI bias) has been studied with respect to organism environment and taxonomy, and subcellular location. It has been proposed that average pI correlates with growth temperature for orthologues [14], and with bacterial growth conditions [15], that pI bias correlates with taxonomy [16], and that pI distribution varies according to compartmentalisation within the Arabidopsis chloroplast [17]. Other work indicates that, for the most part, pI distributions are not correlated with subcellular location or taxonomy [18]. Amongst these varying conclusions, observations reinforced by multiple reports are that: individual protein pIs tend towards less extreme values for longer sequences, as a result of sampling statistics of acidic/basic amino acids [16, 19]; subcellular proteome pIs may give net charge at environmental pH to mitigate against protein aggregation [18, 20]; smaller proteomes
tend to be more basic [15-16]. This last trend is particularly evident for the small proteomes of intracellular parasites, and does not appear to be fully explicable in terms of genome AT bias. Processes suggested to underlie the trend include adaptation to environmental constraints, such as elevated host pH [15], and differences in the rate of accumulation of mutations (higher in intracellular organisms than free-living ones) [16].

3D structures are known for many proteins, and may be modelled for many more [21]. Structure can be used to predict physico-chemical properties, which in turn can aid understanding of function or environmental adaptation, for example comparing proteins from mesophiles and extremophiles [22] or distinguishing Enzyme Commission (EC) classes for enzymes [23]. Charges contribute to protein stability, evident from simple geometric analysis [24]. The role of ionisable groups has been studied extensively, with regard to both protein stability and solubility [25], and in terms of specific functionality, such as proton buffering by hemoglobin [26]. Computational models of charge interactions [27] can be applied across databases, looking for an example at predicted ionisations of amino acids [28]. Varying degrees of model complexity have been introduced, and are assessed through agreement with experiment for properties such as pKas, the pH-dependence of folding energy, and mutational effects. We have found that a relatively simply model for charge interactions captures the properties of surface ionisable groups [29], whereas more detailed accounting of a protein/water interface is required for substantially buried groups [30]. Since the great majority of ionisable groups lie at the surface, the simplified method is appropriate for application to wide-scale structural proteomics, so long as detailed questions are not asked of the more buried, typically catalytic site, ionisable groups. This has been shown in a previous study, where we focussed on comparing isoelectric points predicted from sequence and structure [20]. It was found that predicted pI, averaged over the protein structures in a subcellular compartment, tends to lie away from the subcellular pH, consistent with a role in mitigating against isoelectric point aggregation. The subcellular average of the pH at which proteins are predicted to be most stable appeared to be a closer match to subcellular pH, than was the average pI. This preliminary observation, which is consistent with other computational work finding that the predicted pH of maximal stability can be quite different to the pI [31], is now investigated in detail. We find that the (pH-dependent) maximum in protein stabilisation relates to modulation of histidine pKas by 3D interactions. These residues are largely at the surface and not recognised individually to be of primary functional importance, and
yet their cumulative properties associate with subcellular pH. We discuss the physiological context for this result, including pH homeostasis, pH sensing and stochastic effects.

2.2 Methods

2.2.1 Dataset

Release 55.5 of UniProtKB/Swiss-Prot [45] was searched for annotation according to the following subcellular compartments: nucleus; vacuole; cytoplasm; extracellular; lysosome; chloroplast; mitochondrion; endoplasmic reticulum; peroxisome; Golgi. Entries with uncertain keywords such as similar, potential, probable were omitted. Requiring at least one cross-reference from the Protein Data Bank (PDB) structural database [46] gave 5278 UniProt entries referencing 16311 PDB identifiers. These PDB identifiers were filtered using the PISCES server [47] for X-ray diffraction structures better than 3 Å resolution, a minimum chain length of 30 amino acids and redundancy at 90% sequence identity, yielding 3,713 protein chains. The 90% sequence filter was chosen so that identical chains would be eliminated, but allowing for amino acid variation on a common fold, since the calculated charge interactions will change with such variation. Of the 3,713 chains, we excluded those with extracellular (1,048) and chloroplast (90) annotation, as these locations present a broad pH distribution. A further 534 were annotated with more than one subcellular location and were also excluded, as well as 20 failures in the calculation scheme (for example, due to non-standard residue names). Structure-based predictions of pH-dependent properties were made for the remaining 2,021 protein chains (see Additional File 2-1 from the Additional Files section), roughly double the number compared with previous work [20]. Figure 2.1 summarises the construction of this dataset.
Figure 2.1: Construction of the protein structural dataset
UniProt and the PDB were used to cross-reference subcellular annotation and structure, with filtering for sequence identity and structure quality to give a set of protein chains for calculation. See the Materials and methods section.

2.2.2 Calculations
Continuum models are commonly used for calculating charge interactions in biomolecules. The complexity and computational requirements of these models varies according to the accuracy with which the boundary between solute and solvent is described. In the current work, we require a relatively fast method, enabling calculations not just for many proteins, but also for a
randomised dataset that is generated to evaluate the central hypothesis. A simple Debye-Hückel method is sufficient for these purposes, since most of the ionisable charge proteome is exposed to solvent, with water dominating the solvation response [29]. In earlier work in this area, it was found that the relatively simple Debye-Hückel method gave very similar results to the more computationally demanding Finite Difference Poisson-Boltzmann method [20]. A uniform relative dielectric of 78.4 and an ionic strength of 0.15 Molar were used in calculations of charge interactions. To compute ionisable group pKas [48] from these interactions, Monte Carlo sampling of protonation states was used [49]. Changes in folding energy were derived from the charge difference between folded (F) and unfolded (U) states of a protein ($\Delta Q_{FU}$) [33], with an origin set from the ionisable group contribution to the folding energy calculated at an extreme pH with the reduced sites method [48]. The following model compound pKas were used: sidechains: Asp 4.0; Glu 4.4; His 6.3; Lys 10.4; Arg 12.0; terminal groups: N-terminal 7.5; C-terminal 3.8. Cysteine and tyrosine ionisation has not been considered, since although important in certain catalytic processes, these ionisations are of less interest in a study of global charge properties around neutral pH. The unfolded state is approximated as lacking interactions between ionisable groups. While there are known to be charge interactions in the U state [50], the current work focuses on the relationship between pKas in the folded state (histidine in particular) and subcellular location. Of interest is that ionisable group interactions in the unfolded state appear to be dominated by local sequence neighbours, in part recapitulating the interactions of the folded state [29]. We trialled a simple model for pKa changes in the unfolded state [29] as a mimic for folded state pKas, examining whether the correlations observed with subcellular pH were reproduced. If this had been successful, it would have broadened the study to bypass protein structure in examining proteomes; however this trial failed, with substantially lower correlation observed between calculated properties and subcellular pH, compared with structure-based calculations. The pKa calculation program can be found in Additional File 2-3.

An additional test of the Debye-Hückel-based method for calculating maximal stability was made, following a protocol established in previous work [31]. Briefly, the BRENDA enzyme database [51] was searched for text strings associated with maximal or optimal pH of stability, and these results cross-referenced with PDB entries for the same species and enzyme. The literature references retrieved from BRENDA were checked for data pertaining to a well-defined optimum, rather than a pH-range, and also for confirmation of the optimum in relation to stability.
as opposed to activity. Additional File 2-2 gives information for the 19 enzymes retrieved in this analysis (see Additional Files section). The listed criteria, and in particular the requirement for a precise species match between structure and stability data, leads to a smaller dataset than that reported previously [31]. The calculated root mean square deviation between calculated and experimental pH stability optima is 0.78 for the dataset of 19 enzymes, comparable with the value of 0.72 for the earlier work [31], and supporting the use of Debye-Hückel modelling in this study.

The quantity His/(Acid + Base) was calculated as the number of histidine residues divided by the sum of Asp, Glu, Lys and Arg residues. In many cases, averages of properties were calculated over a set of protein structures annotated with a particular subcellular location, and denoted by <> symbols.

In order to test the importance of histidine sidechains, predictions of pH dependence for proteins were made with the ionisable charges of histidine sidechains removed. Additional tests made use of proteins and their number of histidine sidechain charges, but not their locations. Their positions were randomly assigned to surface atoms (within amino acids of accessible surface area > 5 Å²), ensuring ionisable charges were separated by at least 3.5 Å. This distance constraint applied also to the background acidic and basic groups whose locations were unchanged. One hundred datasets of proteins (each mirroring the 2,021 proteins of the wild-type dataset) were constructed in this way, and pH-dependent features calculated. The extent of these computations required us to look at whether the full Monte Carlo sampling, to obtain pKas, could be circumvented for the properties of interest. It was established that the pKa deviations of histidine sidechains (∆pKa = pKa – pKa[model compound]) could be estimated accurately (in relation to the Monte Carlo sampling) from summation of charge interactions at each ionisable site, at neutral pH (see Results and Discussion).
2.2.3 Subcellular pH

Experimentally-determined pH values for subcellular compartments were collated from various sources in previous work [20]. Their pH values and their original sources are shown in table 2.1. Subcellular pHs were chosen from available publications via online search engine using keywords such as '{subloc} pH', where {subloc} were the names of the 8 different subcellular locations. For nucleus and cytoplasm, their pHs were averaged from 6 measurements using different mammalian cell types. For vacuole, its pH was updated from 6.3 [53] to 5.3 [54] using modern pH measuring technique. pHs for lysosome, peroxisome, ER and Golgi apparatus are from the publications listed in table 2.1 and their values are generally accepted to be within +/- 0.2 pH units. For mitochondria, it is known that pHs are different in the mitochondria matrix (pH 8) and the intermembrane space (pH 7) for a pH gradient, which is needed for ATP synthesis. The ATP synthase couples ATP synthesis to H+ transport into the matrix. We have averaged the two pH values in order to represent its overall trend of basicity. There will be some uncertainty in precise values for individual locations, for example due to compartmentalisation, but the overall trend of more acidic, neutral, or more basic compartments is the key factor.

<table>
<thead>
<tr>
<th>Subcellular Location</th>
<th>Experimental pH</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleus</td>
<td>7.55, 7.79, 7.63, 7.68, 7.77, 7.8</td>
<td>Seksek &amp; Bolard, 1996 [55]</td>
</tr>
<tr>
<td></td>
<td>Average: 7.7</td>
<td></td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>7.24, 7.5, 7.37, 7.20, 7.49, 7.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average: 7.3</td>
<td></td>
</tr>
<tr>
<td>Vacuole</td>
<td>5.3</td>
<td>Kulichikhin et al., 2009 [54]</td>
</tr>
<tr>
<td>Lysosome</td>
<td>4.8</td>
<td>Geisow et al., 1981 [56]</td>
</tr>
<tr>
<td>Peroxisome</td>
<td>8.2</td>
<td>Dansen et al., 2000 [57]</td>
</tr>
<tr>
<td>Endoplasmic Reticulum</td>
<td>7.1</td>
<td>Kim et al. 1998 [58]</td>
</tr>
<tr>
<td>Golgi Apparatus</td>
<td>6.6</td>
<td></td>
</tr>
<tr>
<td>Mitochondria intermembrane space</td>
<td>7.0</td>
<td>Alberts et al., 1994 [59]</td>
</tr>
<tr>
<td>Mitochondria matrix</td>
<td>8.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average: 7.5</td>
<td></td>
</tr>
</tbody>
</table>

*Table 2.1: Subcellular locations and their experimental pHs from various literatures*
2.3 Results and Discussion

2.3.1 pH-dependence of Stability and Subcellular pH

A dataset of protein structures annotated by subcellular location was constructed as described in the Materials and methods section (Figure 2.1). Figure 2.2 illustrates the major ionisation regions for proteins on a schematic plot of the pH-dependence of folding energy ($\Delta G_{FU}$, the difference between the folded (F), and unfolded (U), states). Acidic and basic titrations underpin the generally bimodal pI distributions observed for proteins [32], since the numbers of (Asp + Glu) with acidic pKas, or (Lys + Arg) with basic pKas, normally exceed the number of His, which ionise in the central pH range. While the balance of (Asp + Glu) and (Lys + Arg) mostly determines pI, it follows from the proportionality between $\partial \Delta G_{FU} / \partial \text{pH}$ and $\Delta Q_{FU}$ (the difference in net charge between folded and unfolded forms) [33], that the pH-dependence of stability in the central pH range is determined largely by histidine ionisation (Figure 2.2). Further, this slope (although small when histidine content is low) is generally negative or positive according to whether the protonated state of histidine is stabilised ($\Delta Q_{FU}$ positive) or destabilised ($\Delta Q_{FU}$ negative) in the folded protein. This in turn determines whether the pH value at minimal $\Delta G_{FU}$ ($\text{pH}[\Delta G_{FU}(\text{min})]$, figure 2.2) is towards the acidic or basic titration block, again yielding a generally bimodal distribution. Figure 2.3 shows the $\text{pH}[\Delta G_{FU}(\text{min})]$ distribution for proteins in the 8 different subcellular locations. A bimodal distribution is observed. The general trend is that more proteins with a basic $\text{pH}[\Delta G_{FU}(\text{min})]$ and less proteins with an acidic $\text{pH}[\Delta G_{FU}(\text{min})]$ were found when moving to subcellular locations with more basic pH.

Isoelectric and pH-dependent properties were calculated and examined for correlation with each other and with the measured environmental pH values (Table 2.2). The average across each subcellular compartment of the pH at minimal $\Delta G_{FU}$, denoted $\langle \text{pH}[\Delta G_{FU}(\text{min})] \rangle$ correlates better with subcellular pH than do the analogous averages for pI, $\langle \text{pI}(F) \rangle$ and $\langle \text{pI}(U) \rangle$ (see also Figure 2.4), although none of these properties matches subcellular pH across the entire range. Table 2.2 also shows the correlations when calculations are repeated with histidine ionisations removed. Here, only average isoelectric properties and subcellular pH are correlated, demonstrating that histidine ionisation lies behind the correlation of $\langle \text{pH}[\Delta G_{FU}(\text{min})] \rangle$ with subcellular pH (Figure 2.2). Further, histidine location rather than composition is implicated,
since neither subcellular pH nor $<\text{pH}[\Delta G_{\text{FU}}(\text{min})]>$ correlate with the subcellular averaged ratio of histidine to other charged amino acids, $<\text{His}/(\text{Acid}+\text{Base})>$ (Table 2.2). The nature of pH $[\Delta G_{\text{FU}}(\text{min})]$ versus $\text{His}/(\text{Acid}+\text{Base})$ for individual proteins (not shown) is that more acidic or basic $\text{pH}[\Delta G_{\text{FU}}(\text{min})]$ map to higher $\text{His}/(\text{Acid}+\text{Base})$, while lower His content maps to a relatively underpopulated central zone of $\text{pH}[\Delta G_{\text{FU}}(\text{min})]$, bearing out the schematic indications of figure 2.2.

Figure 2.2: Major ionisation zones in the pH-dependence of protein stability
A schematic diagram of the major ionisation regions in a plot of folding stability ($\Delta G_{\text{FU}}$) versus pH. Properties describing this pH-dependence, in the notional case of no protein unfolding over the pH range, are shown. Cysteine and tyrosine have been omitted from this figure since they are mostly unionised at physiological pH.
Figure 2.3: Normalised distribution of $pH[\Delta G_{FU}(\text{min})]$ for protein chains in the 8 subcellular locations

Bimodal distribution is observed in all compartments: (a) - (c).
<table>
<thead>
<tr>
<th>First property</th>
<th>Second property</th>
<th>$R^2$ (with His)</th>
<th>$R^2$ (without His)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subcellular pH</td>
<td>$&lt;$pI(F)$&gt;$</td>
<td>0.44</td>
<td>0.47</td>
</tr>
<tr>
<td>Subcellular pH</td>
<td>$&lt;$pI(U)$&gt;$</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Subcellular pH</td>
<td>$&lt;$His/(Acid+Base)$&gt;$</td>
<td>0.02</td>
<td>n/a</td>
</tr>
<tr>
<td>Subcellular pH</td>
<td>$&lt;$pH[$\Delta$G$_{FU}$(min)]$&gt;$</td>
<td>0.84</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>$p$(F)</td>
<td>0.70</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>$p$(U)</td>
<td>0.74</td>
<td>0.01</td>
</tr>
<tr>
<td>$&lt;$pH[$\Delta$G$_{FU}$(min)]$&gt;$</td>
<td>$&lt;$His/(Acid+base)$&gt;$</td>
<td>0.02</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 2.2: Correlations between calculated properties and subcellular pH

Squares of correlation coefficients are given. Without His refers to calculations with histidine ionisable groups removed.

---

**Figure 2.4: Subcellular pH and pH-dependent properties**

(a) The average over subcellular compartments of predicted folded form pI, plotted against subcellular pH, with $R^2=0.44$. (b) The pH of maximal folded state stability, averaged over proteins for each subcellular location, is plotted against subcellular pH, with $R^2=0.84$. For both panels, the best fit line is shown and the data are shown in the table underneath.
2.3.2 Histidine pKas and Subcellular Location

Histidine pKa deviations from the model compound value, obtained in the pH-dependence calculation and averaged for each protein, correlate well with <pH[ΔGFU(min)]> (R²=0.99, not shown). In Figure 2.5(a), histidine pKa deviations are plotted against subcellular pH, again with good correlation. These are now averaged per histidine, <ΔpKa[His]>, since the protein-specific condition implicit in pH[ΔGFU(min)] is lost. Figure 2.5(b) shows the compartment-specific <ΔpKa[His]> with the ranges observed, using the 5% and 95% ranked ΔpKa[His] values within each location. Variation across the subcellular averages is much smaller than the variation of histidine ΔpKas. We have investigated previously whether calculations of <pH[ΔGFU(min)]> change substantially upon the inclusion of more a detailed charge interaction scheme, or a model for residual charge interactions in the unfolded state [20]. Although pKas can be perturbed in the U form [34], [29], it was found that the effect of these modelling adjustments on <pH[ΔGFU(min)]> was small [20]. Furthermore, in the current work, we looked at a sequence-based U form model for charge interactions, with nearest neighbour pairs dominating. If this were to yield correlations with subcellular pH, then the analysis would not be restricted to protein structures. However, the U form model proved too simple, yielding relatively poor correlations (e.g. R²=0.29 between <ΔpKa[His]> and subcellular pH), and this line of enquiry was not pursued further.
Figure 2.5: Histidine and subcellular pH

(a) Predicted ΔpKa per histidine imidazole, averaged over histidines in each subcellular location, is plotted against subcellular pH. The line of best fit is drawn ($R^2=0.92$). (b) For each subcellular compartment, the average of calculated ΔpKa per histidine is shown. The 5% and 95% ranked values for ΔpKa in each compartment are shown in the table on the right. LYS, lysosome; VAC, vacuole; GOL, Golgi; ER, endoplasmic reticulum; CYT, cytoplasm; MIT, mitochondrion; NUC, nucleus; PER, peroxisome.
Figure 2.6: Histidine location and subcellular pH
(a) Schematic diagram showing randomisation of histidine locations (green circles in the shaded protein shape), with all proteins in the dataset sampled 100 times in this scheme. Background negative (red) and positive (blue) charges are not moved. (b) Correlations (given as $R^2$) sampled over the 100 randomisations and compared with the actual value ($R^2=0.92$). These correlations are between subcellular pH and computed average $\Delta pK_a$ per histidine (in each subcellular compartment), where the $\Delta pK_a$ calculations follow the simplified scheme described in Methods.
Having established that the predicted and averaged ionisation properties of histidine sidechains are strongly correlated with subcellular environment, but also bearing in mind that average isoelectric points show some correlation, we investigated further the role of histidine positioning in protein structures. Figure 2.6(a) illustrates a scheme in which, for each protein, acid and base charges other than histidine are fixed and the ionisable groups of histidine explore alternate surface locations (see also the Methods section). One hundred passes were made through the entire dataset, randomising the location of histidine sidechain charge for each protein and recalculating ΔpKas. In order to make these computations feasible, estimates of ΔpKa from full Monte Carlo sampling were substituted by summation of acid/base interactions at each histidine site, assuming protonated bases and deprotonated acids. For the set of non-randomised proteins, this procedure gave the same R² (0.92) for <ΔpKa[His]> versus subcellular pH, as did the results of Monte Carlo sampling displayed in figure 2.5(a), and the <ΔpKa[His]> values themselves correlated with R²=0.999 between the two calculations. Figure 2.6(b) shows that the correlation with subcellular pH, for calculations with the real distribution of histidine ionisable groups, exceeds that for the randomisations. The net charge of a protein can influence His ΔpKa, simply due to an environment weighted towards positive or negative charge, i.e. positive overall destabilises histidine protonation leading to negative His ΔpKas, and a negatively charged background stabilises His protonation, giving positive His ΔpKas. Indeed, the net charge has some correlation with subcellular pH, as seen in table 2.2 and figure 2.4(a), whereas figure 2.6 demonstrates that an additional element of correlation with subcellular pH is attributable to histidine location, beyond the net charge background (which remains constant in the randomisations). Reinforcing this conclusion, R² for the correlation between the average of net charge on a protein (excluding histidine), over subcellular location, and subcellular pH is 0.42, substantially less than that for <ΔpKa[His]> and subcellular pH of 0.92. Next we consider the physiological context for our observations.
2.3.3 Relevance of Correlation Between Histidine pKas and Subcellular pH

We have found that protein populations tend towards their most stable, on average, at the pH of the relevant subcellular environment. A couple of notes should be made about these results, which are based on predictions of pH-dependent properties from protein structures. The calculation model is simple (Debye-Hückel), based largely on geometry of the charge network. This works well for groups at the protein surface, with charge-charge interactions dominated by water, which is the vast majority of ionisable groups. In addition, the linear correlation of predicted properties with subcellular pH is good, but the fit between these properties and actual subcellular pH values falls away for the acidic vacuolar and lysosomal compartments (Figure 2.4(b)). Looking back at the distribution of pH$[\Delta G_{\text{FU}}(\text{min})]$ (Figure 2.3), the frequency for acidic pH$[\Delta G_{\text{FU}}(\text{min})]$ is low compared to the frequency of basic pH$[\Delta G_{\text{FU}}(\text{min})]$ in basic compartments. This might be due to a small sample size for proteins in the acidic compartments. Also, there is a restriction imposed by the pKa of ionisable groups, limiting the pH$[\Delta G_{\text{FU}}(\text{min})]$ to be within pH 5 and 9 in general (Figure 2.2). At pH 4 or lower, ASP and GLU will become neutral, while HIS, ARG and LYS are positively charged and are likely to create repulsion between the charge network on the protein surface. An energy penalty will be added to the overall protein stability, giving a high $\Delta G_{\text{FU}}$. A high $\Delta G_{\text{FU}}$ will also be obtained when pH is at 9 or above for the same reason that there are repulsive force between negatively charged ASP and GLU, while other ionisable groups are neutral. However, in pH range between 5 and 8, the charge network on the protein surface are stabilised by opposite charges, making the lowest $\Delta G_{\text{FU}}$ remains in this range.

Taking the basic observation, of predicted maximal stability at subcellular pH, it is important to note that the bimodal distribution of individual protein pH$[\Delta G_{\text{FU}}(\text{min})]$ values means that generally each protein is not most stable at the pH of its surroundings i.e. the observed correlation relates to a sum over proteins in a particular environment. If this correlation were not observed, then in principle the folding free energy of proteins could (on average) be more stabilising at an alternate pH. Thus, with the observed correlation, the unfolded population of proteins is (on average) minimised with respect to subcellular pH. However, pH-dependent changes in $\Delta G_{\text{FU}}$ for each protein, and related alteration in the F/U population, are generally small, but possibly could be significant over the subcellular population of proteins. This is a protein folding perspective on the results.
An alternative view would be to consider protonation, pH buffering and pH homeostasis, which is directly related to protein folding since \( \frac{\partial \Delta G_{FU}}{\partial \text{pH}} \alpha \Delta G_{FU} \) [33]. The regulation of pH is of critical importance [35] and histidine imidazoles are important components of intracellular buffering power [36]. Histidine \( \Delta pK_a \)s underpin the relationships that we observe in the current work. In overall terms, we see that more acidic environments tend towards more acidic \( pH(\Delta G_{FU} \text{(min)}) \), which in turn relates to higher His pKas, more positive His \( \Delta pK_a \)s and relatively stabilised protonated states. For example, His pKas move towards higher values, away from the subcellular pH, for acidic compartments relative to other environments. The general trend is thus to reduce the buffering power associated with His, in each location. However, this movement is small. Taken as an average value per His, the total range of pKa shift between most acidic and most basic environments is about 0.2 pH unit. Although histidine ionisation properties underlie our results, it may be that their direct contributions to proton/pH buffering are not the most important factor.

The reason that small average changes in His pKa give rise to larger changes in \( \langle pH(\Delta G_{FU} \text{(min)}) \rangle \) (displayed schematically in Figure 2.2), is that the stability term includes a difference to the U state, and therefore also to the model compound pKa for His (6.3). Of key importance is \( \Delta pK_a \), determined by charge interactions in the F state, so that if the model compound value changes, the overall result remains. Thus far we have discussed our results in the context of overall protein stability and pH buffering. Next we combine these aspects.

When a protein folds or unfolds it may release or take-up protons. Another way of looking at the correlations we find is that, on average and with the caveat about acidic compartments not falling directly on the line in figure 2.4(b), net proton release or up-take is predicted to be close to zero upon folding or re-folding. However, this need not be the case generally, since metabolic processes leading to net changes in proton concentration are handled by the mechanisms of pH homeostasis [37].

Outside of net changes in protein folding, and without considering intrinsically unstructured proteins [38], a subset of proteins or domains will be unfolded at any given time. It is of interest to estimate the number of histidines associated with this unfolded population. Given a protein density of about 1.35 g cm\(^{-3} \) [39], a volume fraction of around 15% for proteins in the cytoplasm
[40], and an average amino acid molecular weight of 110 daltons, the cytoplasm is approximately 1.8 Molar in protein amino acids. With an estimate of histidine amino acid composition at about 2.3% [41], this gives a histidine Molarity of 0.042. If an average folded state stability is taken at around 25-30 kJ/mole [42], then about 1 in 105 domains will be unfolded, so that an approximate concentration of histidine in the unfolded state is $4 \times 10^{-7}$ Molar. Thus the sub-population of histidine ionisable groups that are transiently in the unfolded state could be larger than the concentration of protons. Of itself this may not be a problem, since transient changes across a compartment will average out. What could be an issue though, is whether changes in the populations of folded and unfolded histidine sites couple to local pH-dependent phenomena. This may be protein folding itself, for example with low numbers of protons at pH 7, fluctuations could impede protein refolding that is associated with proton uptake. Alternatively, the mechanisms of pH-sensing and pH homeostasis could be inappropriately activated and modulated by sufficiently large fluctuations. Whether these processes occur in vivo depends on the details of protein and proton diffusion properties as well as on pH-sensing mechanisms, and their response functions, all of which are unknown at the required level of detail. However, the observed tendency to average protonation changes towards zero for protein folding/unfolding, in each subcellular location, would mitigate against such processes.

2.4 Conclusions

In this work we have asked whether the pH-dependence of organelle sub-proteomes, derived with structure-based predictions, correlates with environmental pH. We find that restrictions imposed by the composition of ionisable groups means that individual proteins have minima in pH-dependence, (the predicted pH at maximal stability), that tend to lie on either side of subcellular pH. Averages over proteins within each subcellular location though reveal a strong correlation with subcellular pH. Investigating further it is found that histidine ionisations and ΔpKas from charge interactions in the folded state underlie this correlation.
While net charge and pI also correlate with subcellular pH, and pI correlates with the pH-dependent properties reported here, the strongest relationship is found between pH-dependence (and histidine ΔpKas) and subcellular pH. Thus, while a net charge relationship with subcellular pH could be understood in terms of solubility and avoiding isoelectric aggregation, there is also the question of what lies behind the observed correlation of predicted pH-dependence and subcellular pH. At face value, it could be simply that folding stability tends towards maximal in each subcellular location. However, it is only the average that gives the correlation, rather than the stability maxima for individual proteins.

We have shown that histidine ionisation underlies the pH-dependence correlation. Further, histidine locations are key since random placement of equivalent numbers of histidines, in preserved charge backgrounds, does not reproduce the strength of correlation. This leads us to consider the proton buffering of histidine, but the differences between subcellular environments (i.e. how much histidine ΔpKas are predicted to move) are relatively small.

Finally, we address the role of histidine ionisation in protein folding/unfolding. The direct implication of our results is a prediction that proton release and proton uptake are balanced in a random subset of folding or unfolding proteins. Presumably such a balance would not be required during net protein synthesis or degradation, since the mechanisms of pH homeostasis regulate proton concentration. We speculate that a balance of proton uptake and release could play a role in guarding against activation of pH homeostatic processes during folding and unfolding fluctuations in a steady state subcellular compartment. This can be examined experimentally, with more detailed characterisation of the dynamics of pH homeostasis mechanisms, and computationally with systems level models. It will also be of interest to study the subcellular and extracellular distribution of protonation changes upon complexation. This extends to protein-protein complexation [43] and to protein-small molecule, for example the Bohr effect in hemoglobin [44].
In addition, the work presented in here could proceed further to find out the locations of histidines that were contributed to their elevated pKas in different subcellular locations. We believed there will be some micro-environmental differences in stability-related and stability-unrelated histidine locations. This further work should be straightforward, where histidine locations with more than 0.5 pKa shift will be studied. A sphere with various radius (e.g. 10 Å or 15 Å) could be used to sample nearby residues from the C-alpha atom of the histidine. Then we should be able to see if there is a link between residue distribution and shifted pKa of histidine, when compare to histidines with unshifted pKas.
2.5 References


34. Oliveberg M, Arcus VL, Fersht AR: pKA values of carboxyl groups in the native and denatured states of barnase: the pKA values of the denatured state are on average 0.4 units lower than those of model compounds. Biochemistry 1995, 34(29):9424-9433.


CHAPTER 3 - Prediction of Protein Solubility Based on Surface Charge Patches

PAPER 2

Pedro Chan¹ (PC), Jim Warwicker¹* (JW)

¹Faculty of Life Sciences, University of Manchester, Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester, M1 7DN, UK

*Corresponding author

JW and PC together designed the study and wrote code for the calculations. The dataset was assembled by PC, and PC performed the calculations and most of the data analysis. JW and PC together wrote the paper.

(Article in preparation)
Abstract

Algorithms to predict protein solubility are gaining increased importance with the growing use of protein molecules as therapeutics, and the attendant requirements for long-term stability at relatively high concentrations. Many parameters have been investigated, for correlation with protein solubility. Those which have persistently shown some promise generally fall into two categories, those that consider polarity and isoelectric point (pI), both in the folded and unfolded forms, and those that come from the amyloid area and consider propensity for β strand-mediated interactions. There have also been studies of polarity and pI as these relate to distinguishing proteins from different subcellular locations and different environmental/phylogenetic niches. In previous work we discovered a systematic variation of predicted histidine pKas with subcellular location, conjecturing that this could relate to some restraint on proton uptake and release in respect of transient unfolding and aggregation [1]. In the current work, this hypothesis was explored using structures of proteins that have been classified experimentally in terms of solubility. Variations of histidine pKa are found to be significantly different between soluble and insoluble proteins and further analysis suggested this could be linked to overall charge properties. Indeed a plausible explanation for the histidine pKa variation with subcellular pH is that charge networks have evolved to incorporate histidine in acidic organelles and exclude them in alkaline environments, based simply on the median pKa of the histidine imidazole group. Returning to properties that correlate with solubility, we found that the size of the largest surface patch that lacks a charge provides a predictor with a maximum accuracy of 0.74, and a ROC plot AUC of 0.76. This result is reminiscent of previous work comparing proteins from hyperthermophiles and mesophiles, where proteins evolved to function at higher temperature tend to have a higher percentage of charges on their protein surface and charges are more closely packed, on average than mesophiles. The higher number of charged groups may be related to enhancing solubility at temperatures where aggregation is caused by non-specific interaction of hydrophobic patches. The number of charges and protein stability did not appear to be correlated, indicating that engineered stability and solubility in proteins, via charges, can be accomplished independently. In practice these properties might be jointly engineered since protein stability contributes to the longer-term resistance to aggregation.
3.1 Background

The isoelectric point (pI) of a protein describes the pH at which the molecule has no net charge. There have now been several studies of how protein pIs are distributed with respect to organism phylogeny, environmental niche, and subcellular environment [2-6]. There exists a basic bimodal distribution of pIs [7-9] (Figure 3.1), originating from the underlying ionisation properties of amino acid (AA) sidechains, with histidine (neutral pH pKa) less numerous than acidic (aspartic and glutamic acids) groups, and basic groups (lysine and arginine). Whilst this is clear-cut, there has been discussion over whether there exist systematic variations in the relative populations of these bimodal peaks, according for example to the biological niche of an organism [3]. A consistent feature of pIs with respect to subcellular location is their tendency, on average, to lie away from the subcellular pH. This feature is assumed to result from a physicochemical requirement for net charge to prevent aggregation in the crowded macromolecular environment [1, 10].

![Figure 3.1: pI distribution](image)

The pI data were calculated from the 2021 protein structures in chapter 2 [1], which shows a basic bimodal distribution.

Whilst pIs can be estimated from protein sequence, and do not generally vary substantially from the folded protein pIs [10]. The variation that does exist reflects charge interaction differences between the folded and unfolded states and gives the pH-dependence of folding energy. In a
study of the predicted pH-dependence of folding stability, it was found that the pH of maximal stability, averaged over proteins in each subcellular environment, correlated with subcellular pH [10]. Thus, there is a property that does correlate with an environmental feature, although it is only calculable from 3D structure, and the correlation is an average across proteins rather than individual proteins. The molecular basis of the correlation is largely the locations and interactions of histidine residues, as represented in their change in pKa upon folding [1]. A simple interpretation of this result is that proteins, on average, tend to be most stable at their environmental pH. However, the pH-dependence of folding energy around neutral pH (e.g. cytoplasm) is relatively small, a few kJ/mole. An alternative proposal has been made that there might be some restraint on proton exchange upon unfolding, since too large an exchange could alter the kinetics of refolding after transient partial unfolding [1]. This amounts to a solubility/aggregation hypothesis, based on the role of histidines in the observed correlation.

As with all process in nature, the folding process of proteins requires energy and must obey the laws of thermodynamics. A protein always folds, so that it achieves the lowest possible energy [44]. It is believed that the native folded three-dimensional conformation of a protein is its lowest free energy state, or one of its lowest [45]. There are two theories on this topic. One proposed by Anfinsen, that the protein spontaneously assumes the conformation of lowest energy for a given environment (pH, temperature, ionic strength, etc.), meaning the correct fold of the protein depends upon free energy considerations without the need to consider the AA combinations. Others suggested that the correct fold of the protein need not to be the absolute lowest free energy. The protein is achieving the kinetically available lowest free energy state because there might be high energy barrier for other alternative states. By given enough time, the protein could reach a lower energy state, however the transformation process would be too slow that this could be assumed to be non-existent. Therefore, the practical conclusions are the same from both theories. Under normal circumstances, protein folds into its native conformation of the lowest available free energy.

However, not all proteins achieve their correct conformations spontaneously. Some proteins require other proteins called chaperones to assist their folding process after leaving the ribosomes. Many chaperones are heat shock proteins (HSP), they repair misfolded proteins caused by elevated temperatures or other cellular stresses [46]. Chaperones are also known to
prevent protein aggregation in the crowded cytoplasm and to accelerate protein folding since folded proteins occupied less volume than unfolded chains. Rhodanese, a mitochondrial enzyme that is known to require five different chaperone-type proteins to assist its folding in two steps. Rhodanese binds to DnaK and DnaJ for partial folding, then GrpE catalyses its transfer to GroEL and GroES, where the process ends with a final folded rhodanese. The discovery of chaperones leads us to think about our previous observation that on average, proteins tend to be most stable at their environmental pH [1]. Our pH-dependent features were calculated from protein structures, which came from crystallised proteins of Protein Data Bank. We imagined if the majority of proteins in our dataset do require the assistance by chaperones for correct folding, i.e. their crystallised structures were not in their lowest free energy states, then we would expect a higher correlation (proteins will be more stable in their own compartmental pHs) to be existed in reality. This is because the protein may have folded into a higher free energy state after being synthesised, the assistance of the chaperones isolated this protein from its surroundings, which may provide a suitable environment for it to fold into a correct conformation (by getting over some high energy barriers). Once the protein is stable, it would require more energy to unfold completely. And it should be more adaptable to the pH than it first leaves the ribosome.

We are aware that the data we collected might be biased to naturally self-stable and crystallisation-friendly proteins, and cannot include proteins with no known structures, which may have higher flexibility, various conformations or higher charge distributions. We imagined if such proteins were included in our dataset, there will be minimal effect to our results if their population in each subcellular compartments is small.

With regard to computational work on protein solubility, there has been more sequence-based work, than from 3D structure. A key question is how solubility is defined in developing test sets, for example, making the distinction between proteins that form inclusion bodies, and those that do not [11], to defining all proteins with known structures as soluble [12], and therefore defining all PDB proteins as soluble. Several studies use this classic definition of a soluble protein, folded in aqueous phase, whereas for examining the correlation of pH-dependence and other properties with solubility for the current study, the distinction between proteins that are easier to express in soluble form versus those in inclusion bodies, is more relevant. In this context, the
subcellular environment of choice is the *E. coli* cytoplasm, and the ideal dataset would be inclusion body (IB) forming and non-IB forming proteins, under constant conditions of folding, and any fusions for the expression construct.

The original work in this area [11], uses the IB-forming/non-IB forming distinction, and was based on sequences, but with varying expression conditions. The two calculated properties correlating best with IB formation are charge average (more net charge, less IB), and turn forming residue fraction (more gives IB, perhaps due to slow folding, e.g. proline). Studies from Idicula-Thomas and Balaji [13-16], also uses the IB/non-IB distinction, with both sequence and structure, although the available structural information in these sets is relatively small. The authors report various sequence and non-sequence based correlations with IB/non-IB, including β-forming AA (IB), and thermostability (non-IB). Other computational studies tend to use the PDB=soluble definition, and machine learning techniques are effective, but often are not mined to give physico-chemical interpretation. For example, studies by Smialowski *et al.* (2007) [17] and Magnan *et al.* (2009) [12] consider all proteins with known 3D structures as soluble by definition, and use various data sources, including the TargetDB database [18] of progress in expression for structural genomics, and apply machine-learning techniques.

Work from Tartaglia and co-workers [19, 20] looks at the relationship between mRNA levels and protein solubility in *E. coli*. Proteins with sequence more prone to aggregation are generally expressed at lower levels, where AA polarity is used to indicate aggregation potential, i.e. proteins with a more non-polar sequence have lower mRNA levels. Datasets used were from the CyberCell Database [21] of E.coli proteins, looking at the 50 proteins with highest and 50 with lowest expression, and the lowest and highest expression subsets from an expression library of human proteins in *E. coli* [22]. The solubility dataset from Büssow *et al.* (2004) [22] gives scales of 0-3 for each of expression strength and soluble expression strength. The REFOLD Database [23] annotates proteins as soluble or insoluble, but in practice relates generally to proteins that have been through IBs, rather than making a distinction of entry into IBs or not.

There are a number of aggregation prediction schemes based on the experimental observation that many proteins can be induced to adopt an amyloid, β-rich conformation, under certain
(often denaturing) conditions [24]. These include TANGO [25-27]. Such schemes can include many factors, but generally, the $\beta$-forming propensity for segments of AA sequence is an important element. The surface polarity approach has been adopted in the redesigning of protein surface to improve solubility [29], with a key part being the introduction of groups to break up non-polar patches in 3D structures. This idea is reminiscent of the discovery that there are more charges on protein surfaces in hyperthermophiles, on average, than those from mesophiles [28]. In that case, it is assumed that higher temperature increases the entropy of the proteins in hyperthermophiles, thus may increases hydrophobic interactions between proteins via their surface non-charged patches. Therefore, the presence of a higher number of charges help breaking up the non-charged surfaces. This is supported by the observation in hyperthermophiles proteins, where smaller non-charged patches are found compared to mesophiles [28]. In addition, the high number of charges found in hyperthermophiles is not well correlated with, the contribution to stability of their proteins. The newer work [43] includes a molecular dynamics analysis to add protein flexibility, and its influence on surface properties. One interesting general question is the extent to which analysis of dynamic properties, fluctuations, and partial unfolding, are required to account for protein solubility.

Available methods for predicting protein solubility have been reviewed recently [30]. A significant problem with assessing these methods remains the available experimental data. An important study, in which a high throughput system was developed for classification of solubility, based around lack of inclusion body formation, in *E. coli* [31]. From these data, it was assessed that factors correlating to some extent with solubility include charge and structural class, whilst algorithms based largely on propensity to form $\beta$-structure/amyloid, performed less well in correlations with solubility. In the current work, histidine ionisation [1], and charge patches [28], are examined in various subsets of solubility test data, mapped onto 3D structure. An earlier report of differential solubility properties according to structural class [31] is followed up, with the suggestion that combination of patch properties and amyloid-forming propensity could lead to a more effective predictor than that with charge patches alone (which gives an AUC of 0.76 and accuracy of 0.74). Although a one-dimensional, sequence-based, charge patch predictor can also be constructed, it is likely that in many cases of designing for solubility and long-term resistance to aggregation e.g. with antibodies, structures or 3D models will be available.
3.2 Methods

3.2.1 Datasets

Six protein sets were used in this study, representing soluble and inclusion bodies (insoluble) sets from three different sources. Since we were interested in the features derived from protein structures, PDB (Protein Data Bank, [32]) files were retrieved by cross-referencing using the Uniprot Knowledgebase (Release 15.2) [33] where the protein identifiers (IDs) obtained did not directly indicate a PDB structure. The nature of the proteins is indicated in the dataset name, followed by 1 or 2 to represent soluble and insoluble, respectively. Table 3.1 summarises the filtering steps for each set, which are also set out in more detail in the following paragraphs. Additional File 3-1 contains the original tables of protein IDs, which our datasets were derived from.

Datasets I1 and I2 were obtained from table S6 of Idicula-Thomas and Balaji [15], which consists of 40 soluble proteins (SP) and 130 inclusion body proteins (IB) from their datasets I, S and T. The sets consist of proteins over-expressed in *E. coli* were originally obtained by searching a combination of keywords: soluble, inclusion bodies, *E. coli*, and over-expression in PubMed. Their protein IDs were a mixture of NCBI accession number and EMBL IDs and PDB IDs, thus a cross-referencing step was performed. These PDB IDs were then filtered by the PISCES sequence culling server [34] to remove sequences with more than 90% similarity at per chain level, obtaining 19 and 18 protein chains in I1 and I2 respectively.

Datasets B1 and B2 were obtained from additional file 1 associated with Büssow *et al.* (2004) [22], which consists of 1287 Ensembl human mRNA transcript IDs. They were expressed in *E. coli* and their expression strength (ES) and soluble expression strength (SES) were measured on a scale between 0 (low) and 3 (high). Ensembl IDs were extracted into B1 if ES=3 and SES=3 and B2 if ES=3 and SES=0. This gave 241 and 102 Ensembl IDs in B1 and B2 respectively. Following PDB cross-referencing and sequence culling by PISCES at 90% cut-off, 45 and 14 protein chains remained in B1 and B2 respectively.
Datasets N1 and N2 were obtained from the data file downloaded from eSOL (the E.coli solubility database, URL: http://tp-esol.genes.nig.ac.jp/index.php?lang=en), associated with the study of Niwa et al. (2009) [31]. E. coli proteins were expressed in the PURE cell-free and chaperone-free system [35, 36]. The data file contained information such as the gene locus IDs, solubility percentage, subcellular location and cross-reference PDB IDs. Our definition of SP and IB for the N1 and N2 sets followed the description of Niwa et al. (2009) [31]. The entry is considered to be soluble if the solubility is more than or equal to 70%, whereas less than or equal to 30% is considered as insoluble. Entries with the specified solubility and an associated PDB ID were separated out into N1 and N2, this gave 171 and 78 entries respectively. In addition to culling at the 90% sequence identity level, it was also necessary to remove membrane proteins in the high throughput data from this study, leaving finally 111 protein chains in N1 and 56 protein chains in N2.

Our requirement for cross-referencing to the PDB, to allow structure-based calculations, gives a big reduction in protein numbers. Some calculations were carried out on all the datasets, and then further work on the N1 and N2 datasets alone, where the higher number of proteins gives a more reliable analysis, particular when the datasets are further divided into the structural class (with reference to SCOP fold classifications [37]).
Table 3.1: Soluble and Insoluble protein dataset selection and filtering.

3.2.2 Calculations

3.2.2.1 Histidine Sidechain Ionisation

In previous work (Chapter 2), we found a correlation between the subcellular pH of an environment and the “average of predicted histidine ΔpKas in that environment” (<ΔpKa[HIS]> [1]. This correlation is clear when averaged over proteins, and is dependent on histidine location rather than simply the number of histidine residues in a protein. One issue that was raised in our previous work is whether histidine ionisation could play a general role in protein folding/unfolding, beyond the occasional specific function of a key interaction. It was speculated that a balance of proton uptake and release upon unfolding could mitigate against fluctuation phenomenon leading to enhanced local pH changes and further unfolding. Effectively this hypothesis is related to protein solubility, implying that histidine ionisation properties are
factoring into solubility via localised unfolding. Since histidine ionisation properties and folded state stability around neutral pH are closely linked, it is also possible that proteins have simply evolved to be most stable at their environmental pH. The problem with the stability hypothesis is that the energy differences (pH-dependence of folded state stability) around neutral pH are mostly quite small.

In order to examine whether ΔpKa of histidines (ΔpKa[HIS]) is related to solubility effects, this property was calculated for the structurally annotated datasets of soluble and insoluble proteins, as described previously (Chapter 2) [1]. Briefly, Debye-Hückel (DH) charge interaction calculations were carried out for ionisable groups, with a uniform relative dielectric value of 78.4 and an ionic strength of 0.15 Molar. The DH model gives similar results to the Finite Difference Poisson-Boltzmann method, for ionisable groups at the water exposed surface of a protein, which corresponds to the large majority of such groups. It is also computationally less demanding [38]. These charge interactions are then used in a Monte Carlo sampling of protonation states [39], in a framework for pKa calculations [40]. Model compound pKas were used: Asp 4.0; Glu 4.4; His 6.3; Lys 10.4; Arg 12.0; N-terminal 7.5; C-terminal 3.8. Cysteine and tyrosine were not included because their effects are small around neutral pH.

ΔpKa[HIS] is defined as the pKa of histidine in the folded state minus its model compound pKa of 6.3. It is organised in three different ways to become a feature to represent individual proteins: (1) The sum of all ΔpKa[HIS] in a protein is denoted as SumΔpKa[HIS]; (2) The sum of all ΔpKa[HIS] in a protein divided by the number of histidines is denoted as “SumΔpKa[HIS] (by HIS)” and this is identical to the average value of all ΔpKa[HIS] in the protein; (3) The sum of all ΔpKa[HIS] in a protein divided by the number of AAs is denoted as “SumΔpKa[HIS] (by AAs)”. Figure 3.2 shows the SumΔpKa[HIS] in a schematic view.
Born solvation energy, background charges and interaction energy between ionisable groups are presented in the protein structure, which affect pKa values of the ionisable groups such as HIS shown above. 

<table>
<thead>
<tr>
<th>HIS pKa in Protein Structure</th>
<th>Model Compound pKa</th>
<th>ΔpKa (difference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3</td>
<td>6.3</td>
<td>1.0</td>
</tr>
<tr>
<td>6.5</td>
<td>6.3</td>
<td>0.2</td>
</tr>
<tr>
<td>6.0</td>
<td>6.3</td>
<td>-0.3</td>
</tr>
<tr>
<td>5.6</td>
<td>6.3</td>
<td>-0.7</td>
</tr>
</tbody>
</table>

SumΔpKa[HIS] = 0.2

Figure 3.2: Schematic diagram to demonstrate calculation of ΔpKa[HIS]
This represents a general scheme, applicable to continuum electrostatics calculations. For the DH model, relative dielectric is uniformly at the value for water, 78.4, and then the component due to Born solvation energy for charge placement in the protein is zero.

3.2.2.2 Largest Non-charged Surface Patch
The method for determining the size of surface patches that do not contain a group that is likely to be charged around neutral pH (Asp, Glu, Lys, Arg, His, N-terminus, C-terminus), was introduced by Greaves and Warwicker (2007) [28]. Briefly, a grid-based shell framework, developed previously to study enzyme active sites [41], was used. The shell is a continuous surface around the protein inscribed onto a grid. Spheres are centred on each of the charged groups, and all grid shell points that lie within a sphere are assigned as charged, and points that do not lie within any sphere are assigned as uncharged. Charged and uncharged patches are then contoured on this shell. At low values of charge sphere radius, uncharged regions connect to form a single large patch, and at larger sphere radius, the charged patches connect, isolating uncharged patches. Then the sizes of all patches are recorded for each protein, and various properties compared between the datasets of soluble and insoluble proteins. The largest uncharged surface patch (Lg2DnonQpatch), was recorded using various sizes of charged
sphere radii (6.0, 7.0 and 8.0 Å). The same is also performed for the ratio of summed uncharged and charged patch areas (Sum of Lg2DnonQpatch Ratio).

3.2.2.3 Patch Modification for Compact Conformations

Following a report that solubility properties differ between structural classes [31], with α-helical proteins having a relatively larger peak of insoluble proteins than β-proteins, an adjustment to the patch analysis was devised. It was hypothesised that α-helices at a protein surface may be more prone to transient unfolding, since the relatively compact helical conformation could, in some cases, unwind without necessarily disrupting neighbouring protein fold. By contrast, the extended β-structure would not be able to locally and transiently unfold to such a degree. It was also hypothesised that such unfolding events could lead to insolubility through aggregation around exposed non-polar segments.

The optional modification is a halving of the sphere drawn around a charged group, if the Cα of that charged residue lies in a segmented determined to be relatively compact structure. This, in turn, is judged from the Cα-Cα distances between this residue and those 4 residues ahead and 4 residues behind. If the average of these distances is < 7.0 Å, then the region is judged as compact. Note that this distance for an α-helix is about 6.2 Å, so that a charged group within an α-helix (a helical turn or more from each helix end point) would be determined as compact.

3.2.2.4 Longest Non-charged Sequence

In addition to the surface patch properties (i.e. a 2D patch, but in 3D), sequence-based (1D) properties were also calculated, but only for proteins in the Niwa datasets listed in table 3.1, because they are greater in protein numbers. The longest non-charged sequence for a protein is calculated using a PDB chain sequence that is read from the SEQRES line of the PDB file. These three-letter AA codes are translated into the one-letter AA codes. Residues of Asp, Glu, Lys and Arg (omitting His in this case) are then assigned with ‘1’ while the rest are assigned with ‘0’. Short segments of ‘0’s are sorted in descending order, and the number of AAs in the longest segment of ‘0’s recorded (Lg1DnonQsequence). The charge ratio (the sum of Asp, Glu, Lys and Arg divided by the protein chain length) for each protein is also calculated.
3.2.2.5 Mann-Whitney Test

Tests of statistical significance for the differences observed in the features between the datasets were carried out using the non-parametric Mann-Whitney test in the SPSS (Statistical Package for the Social Sciences) software version 16 or using the wilcox.test function in R, a software for statistical computing. Both softwares produced identical Mann-Whitney test results.

3.2.2.6 Receiver Operating Characteristic (ROC) Curve

In addition to the statistical tests of similarity, for some cases, the ability of single or combined features to discriminate between SP and IB datasets was tested with ROC plots. A ROC plot gives the true positive rate (TPR, fraction of true positives out of the positive set) versus false positive rate (FPR, fraction of false positives from the negative set), for a binary classifier. Where TP and FP are true positives and false positives, and TN and FN are true negatives and false negatives, TPR = TP / (TP + FN) and FPR = FP / (FP + TN). The accuracy, a single measure of overall performance for the binary predictor, is recorded as (TPR + 1-FPR) / 2. All values, TPR, FPR, accuracy, scale between 0 and 1, with TPR and accuracy ideally at 1, and FPR ideally at 0. An example of a binary predictor is to introduce a threshold for largest uncharged patch size, and classify positives and negatives relative to that threshold, where larger uncharged patches should relate to insolubility. In the ROC plot, a family of predictions is made by varying this threshold, and the area drawn out under the curve (AUC) is a another measure of the ability of a property as a predictive feature, where 0.5 is random and 1 is ideal.
3.3 Results and Discussion

3.3.1 Histidine Sidechain Ionisation and Solubility

Results for the $\Delta pK_a$ of histidines analysis are showed in table 3.2, which includes variations according to what we have described previously. The smaller datasets from Idicula (referring Idicula-Thomas) and Büssow show no significant difference, i.e. these properties do not effectively separate soluble and insoluble proteins in these datasets, whilst the P-values are consistently below 0.05 for the Niwa datasets (except for Sum$\Delta pK_a$[HIS], which is just marginally over 0.05), indicating there is a significant difference, especially when Sum$\Delta pK_a$[HIS] values were divided by the number of HIS or AAs. From the Mann-Whitney tests, it seems that the relationship between protein solubility and these $\Delta pK_a$[HIS] derived features is not large, otherwise a significant difference could be observed in the small datasets. The combined datasets (IBN1 and IBN2) gave lower P-values (although still > 0.05) than Idicula and Büssow sets, which did benefit from the large samples of Niwa sets. The outcome of the P-values depends on how the raw data were distributed. Figure 3.3 shows individual $\Delta pK_a$[HIS] values in each dataset. For Idicula and Büssow sets, the data range is quite narrow (between -1 and 1) compared to the Niwa set, which has a wider range (between -1.5 and 2) of $\Delta pK_a$[HIS] values. Therefore it is not surprised to get no relations for Idicula and Büssow sets. Additional File 3-2 contains all raw result data for the Niwa sets.

<table>
<thead>
<tr>
<th>Features</th>
<th>I1 (19)</th>
<th>B1 (45)</th>
<th>N1 (111)</th>
<th>IBN1 (175)</th>
<th>I2 (18)</th>
<th>B2 (14)</th>
<th>N2 (56)</th>
<th>IBN2 (88)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta pK_a$[HIS]</td>
<td>0.484</td>
<td>0.5207</td>
<td>1.59E-06</td>
<td>6.82E-05</td>
<td>0.412</td>
<td>0.943</td>
<td>0.051</td>
<td>0.25</td>
</tr>
<tr>
<td>Sum$\Delta pK_a$[HIS]</td>
<td>0.504</td>
<td>0.831</td>
<td>0.0002</td>
<td>0.052</td>
<td>0.386</td>
<td>0.943</td>
<td>0.002</td>
<td>0.083</td>
</tr>
</tbody>
</table>

Table 3.2: Mann-Whitney P-value results for soluble and insoluble protein datasets, using calculated $\Delta pK_a$[HIS] values and their derived features.
**Figure 3.3:** Raw data for $\Delta pK_a[\text{HIS}]$ in each dataset (I1, I2, B1, B2, N1, N2)

**Figure 3.4:** Distribution for $\Delta pK_a[\text{HIS}]$ of N1 and N2 datasets

The distribution has been normalised to percentage of proteins in that set.
Also, we obtained a very low P-value for ΔpKa[HIS] between N1 and N2, indicating there is a statistical difference on ΔpKa[HIS] between soluble and insoluble sets. A normal distribution of ΔpKa[HIS] is observed in N1 and N2 (Figure 3.4), with a higher percentage of positive ΔpKa[HIS] used in soluble set and also a higher percentage of negative ΔpKa[HIS] use in the insoluble set. This observation is more clearer when plotting the distribution of the “SumΔpKa[HIS] (by HIS)” values (also equivalent to “the average of all the ΔpKa[HIS] in a protein”). They are later found to be related to other charge properties, since the overall charge network on a protein surface dictates histidine interactions. Thus, more soluble proteins have, on average, more charge, which gives the increased positive ΔpKas seen in figure 3.5. The insoluble proteins, on the other hand, relate directly to the size of large non-charged patches, which are studied in a following section. The next section addresses the original observation of adaptation of protein pH-dependence to subcellular pH.

Figure 3.5: “SumΔpKa[HIS] (by HIS)” distribution for N1 and N2 datasets
The distributions have been normalised to percentage of proteins in that set.
3.3.2 Histidine Sidechain Ionisation and Subcellular pH

Investigation of whether the ΔpKa[HIS] derived features can separate soluble and insoluble proteins originated from our hypothesis (Chapter 2) [1], that histidine ionisation could link transient unfolding to longer-term aggregation, through local pH fluctuations. In practice, it appears that any association of ΔpKa[HIS] with solubility prediction may be correlated primarily with other charge properties. To make it easy for the readers to follow, the observation of <ΔpKa [HIS]> correlation with subcellular pH (Chapter 2) [1] is revisited. A plausible explanation presents itself, that is rather more simple than relatively complex arguments about protonation and local pH fluctuations. First, one considers that the model compound pKa for a histidine sidechain is 6.3, and that pKas for surface (solvent-exposed) groups generally do not move more than 1 or 2 pH units at most from the model compound value, up or down. Second, taking into account the subcellular pH values, either more acidic than this ‘median’ pKa (vacuole, lysosome), or increasingly on the alkaline side, three regimes can be inferred. At pH < 6.3, the imidazole groups tend to be protonated and thus available to form part of the stabilising charge network of a protein [42], and accordingly will be themselves stabilised in the charged form, with positive ΔpKas (to the left in figure 3.6). At pH > 7.3, since ΔpKas for surface groups are generally < 1, there is only a small chance of a (surface) histidine participating in a charged group network, and thus the observed ΔpKa is close to zero (to the right in figure 3.6). In between these schemes, for intermediate pH values (6.3 to 7.3), there is the opportunity for imidazole group protonation and participation in charge networks (intermediate region in figure 3.6). This hypothesis is consistent with the observed correlation of <ΔpKa[HIS]> with subcellular pH, and with the known properties of histidine ionisation, and describes precisely an environmental adaptation, in this case of protein surface charge networks to subcellular pH.

The average ΔpKa[HIS] (red line in figure 3.6) goes up rather than down at the acidic pH reflects the frequency of the ΔpKa[HIS] data. Histidine sidechain has a model compound pKa of 6.3. The majority of pKas of HIS in vacuolar and lysosomal proteins are found to have more basic pKas, thus their ΔpKas are positives, for example: 7.0 (pKa in protein) - 6.3 (model compound) = 0.7. The protonation of the imidazole group of HIS allows the charge to form part of the stabilising charge network of a protein [42]. The elevated pKa has explained this behaviour since a basic pKa means a stronger binding of the proton, i.e. weaker for the proton
to dissociate. Thus the protonated histidine is itself stabilised and become more stable in the charge network.

**Figure 3.6: Schematic diagram showing how histidine sidechain charge interactions may adapt to subcellular pH**

At more acidic pH, histidine (H) is involved in a network of (on average) favourable +/- interactions, and at alkaline pH, being too far from its intrinsic pKa of 6.3, it remains unprotonated and is not incorporated into the charge networks. Intermediate pH value give mixed behaviour and resultant ΔpKas between the more acidic and more alkaline regimes.
3.3.3 Non-Charged Patch Properties

Charge patch features were calculated for the N1 and N2 datasets only due to their larger size and this potential discriminative power, relative to the I1, I2, and B1, B2 datasets. The P-values for patch and other features, in N1 and N2 discrimination, are shown in table 3.3. Ratios of summed charged and non-charged patch areas do not yield significant P-values (< 0.05), whereas those based in the single, largest non-charged patch do show significantly different distributions between the soluble and insoluble proteins. In addition, a sequence-based analysis, effectively a 1D non-charged ‘patch’ (Lg1DnonQsequence) also gives a significant result. To test whether the results from Lg1DnonQsequence gave duplicate result as effect as Lg2DnonQpatch (6.0Å), a scatter plot is generated (figure 3.7) and the correlation between the two features (for both N1 and N2) is shown. It seems that although N1 (soluble set) has more longer (> 20 AAas) non-charged sequences than N2 (insoluble set), the lengths of these sequences do not directly proportional to the size of those non-charged 2D patch; their data are scattered and has a low correlation. However, a stronger correlation is found in the N2 set between the two features, suggesting that the longest non-charged sequence and the largest non-charged patch from the protein may derived from the same residues. Further, the fractional content of AAs expected to be fully ionised at neutral pH, also gives distinguishable distributions for proteins in the N1 and N2 sets (Table 3.3).

<table>
<thead>
<tr>
<th>Features calculated for “N1 and N2”</th>
<th>Mann-Whitney P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lg2DnonQpatch (6.0 Å)</td>
<td>0.0001</td>
</tr>
<tr>
<td>Lg2DnonQpatch (8.0 Å)</td>
<td>0.005</td>
</tr>
<tr>
<td>Sum of Lg2DnonQpatch Ratio (6.0 Å)</td>
<td>0.408</td>
</tr>
<tr>
<td>Sum of Lg2DnonQpatch Ratio (8.0 Å)</td>
<td>0.433</td>
</tr>
<tr>
<td>Lg2DnonQpatch / (Lg2DnonQpatch + Sum of Lg2DnonQpatch) Ratio (6.0 Å)</td>
<td>0.121</td>
</tr>
<tr>
<td>Lg1DnonQsequence</td>
<td>0.0003</td>
</tr>
<tr>
<td>Charge Ratio (No. of DERK / No. of AA)</td>
<td>0.007</td>
</tr>
</tbody>
</table>

Table 3.3: Mann-Whitney test P-values of features calculated for comparison of the N1 and N2 datasets.
Figure 3.7: Scatter plot with the square of correlation coefficients, comparing 2D against 1D largest non-charged patches, for N1 and N2 datasets

Patch values in 2D are numbers of points (on a 2 Å grid spacing), and in 1D are simply number of AAs (without a charged group).
3.3.4 ROC Plot Analysis of 2D Non-charged Patches, and Separation by Structural Class

Previous work [28] found that the largest (2D) non-charged patch size is an important distinguishing feature between proteins from mesophiles, and those from hyperthermophiles, an effect hypothesised to be related to solubility. Given that a similar distinction is found here, with less soluble proteins tending to have larger non-charged patches (Figure 3.8), a ROC plot analysis was performed, in this case with a charged sphere radius at 7.0 Å. The aim here is to examine the effectiveness of 2D non-charged patch as a predictor for solubility, as the threshold patch size for solubility is varied. The ROC plot shown in figure 3.9, giving FPR against TPR, has an area under curve (AUC) of 0.76, and a maximum accuracy of 0.74 with threshold patch size = 1020. AUC is a measure of how well the parameter, i.e. Lg2DnonQpatch, can distinguish between soluble proteins and inclusion bodies.

The best accuracy was obtained by (1) binning the Lg2DnonQpatch from each protein into bin size of 20 from 20 to 4000 according to soluble or insoluble protein sets. (2) Normalised cumulative frequency was obtained. (3) TPR, FPR and ACC is calculated using the equations mentioned in section 3.2.2.6. (4) The highest value in ACC becomes the best accuracy and the TPR and FPR on the same row become the sensitivity and 1 - specificity respectively. (5) The corresponding bin size become the threshold patch size for the best separation.
Figure 3.8: Distributions of N1 (Soluble) and N2 (Insoluble) datasets, with respect to the largest 2D non-charged patch size

Figure 3.9: ROC plot for the separation of soluble and insoluble proteins from the N1 and N2 datasets

The threshold number of points in the largest 2D non-charged patch is varied. AUC is 0.76, and maximal accuracy of 0.74 for a points threshold of 1020.
Following a reported higher aggregation propensity in all α-helical structural classes [31], the analogous ROC plot analysis was made when the conformational adjustment given in the Methods section was applied, namely that charge spheres for charged groups in relatively compact conformations, were reduced to one half of the standard value to 3.5 Å. A relatively compact conformation is assigned where average distance from a central Cα to neighbouring Cα atoms four AAs upstream and downstream, is < 7 Å. There is very little change in the ROC plot statistics (not shown). Our rationale had been that α-helical regions might have the flexibility to locally unfold, exposing non-polar surface to aggregation processes, in preference to β-strand regions. If this were the case, and if the simple analysis for compact conformation were effective, then it might be expected that solubility prediction would improve. Since this is not the case, it was decided to examine more closely the structural classes (Figure 3.10).

Figure 3.10: Soluble and insoluble proteins from the N1 and N2 datasets are plotted according to overall SCOP domain structural class
The horizontal axis of each plot is the number of points in the largest non-charged patches used for the ROC plot analysis of figure 6, and the equivalent value, but with the conformational adjustment, is plotted for each protein on the vertical axis. The y=x line is plotted in green.
It is apparent (Figure 3.10) that the conformational adjustment has relatively little affect overall in the non-charged patch points. The single largest change (deviation from $y=x$ line) is for an $\alpha$-helical protein, as expected, but the small differences in general explain, in part, why the ROC plot properties do not change substantially upon application of the conformational adjustment. It would be beneficial to proceed to a more drastic modification if we know whether a high number of main chain compact conformations was the primary cause for the worse performance in the current non-charged patch analysis. Figure 3.10 is informative in this regard, bearing in mind that the points threshold for best performance in terms of accuracy, is 1020 points. It is seen that soluble proteins are mostly lower than this threshold, particularly for the all $\alpha$ and $\alpha+\beta$ classes, indicating that non-charged patch is a constraint on solubility, and consistent with a TPR close to 1 at peak accuracy. It is also clear that there exist insoluble proteins with non-charged patches less than the threshold, and this is most the case for the $\alpha/\beta$ structural class. This contributes to a relatively poor FPR at peak accuracy, and this should be the subject for model improvement. The aim of conformational adjustment was to increase non-charged patch sizes, mainly focusing on the compactness of helical structures within the protein. From figure 3.10, it appears that beta sheets could have been included in the conformation adjustment.

3.4 Conclusions

This study of protein solubility was initiated with the hypothesis that the (on average) adaptation of histidine ionisation to subcellular environment (Chapter 2) [1], could be related to protein solubility, through transient unfolding and local pH fluctuation pathways that have not been specified in detail. Data were collated for protein solubility to address this issue, and were significantly enhanced with the high throughput data available from a recent study in *E. coli* [31]. Some separation of soluble and insoluble proteins was seen with the calculated $\Delta pK_a[HIS]$ property, but this is also seen for other charge properties, to which histidine ionisation will be sensitive. Our preferred explanation for the adaptation of histidine ionisation to subcellular environment is summarised in figure 3.6. Simply put, histidine imidazole groups are either incorporated into charge interaction networks, or not, according to whether they are themselves likely to be charged at the environmental (subcellular) pH.
Having embarked upon the solubility study, we were able to call on previous work from the group comparing charge properties in proteins from mesophiles and hyperthermophiles [28]. This work had shown that non-charged patches on protein surfaces differed significantly between these sets of proteins, and it was hypothesised that this could result from solubility effects, since the hydrophobic effect would be higher at increased temperature. Applying the same methodology to the *E. coli* soluble and insoluble protein datasets [31], substantial separation was found, generating a ROC plot for binary classification with AUC = 0.76 and a maximum accuracy of 0.74. The resultant prediction scheme for protein solubility, which is based on 3D structure, is relatively simple and easy to apply. It is also consistent with previous work, sometimes with relatively complex algorithms, that identify surface polarity as a key factor in determining protein solubility [29].

There are at least three directions to proceed for improving this solubility prediction scheme. First, whether sequence-based prediction can substitute for structure-based prediction. Preliminary work in here shows some correlation (Figure 3.7) between 1D and 2D patches, but it is not clear at present that sequence alone will be as effective as structure. However, there is plenty of scope for using comparative models of protein structure, where an experimental structure is not available, for example with antibodies, which are of increasing importance to the biopharmaceutical industry. The second area is related to the first, introducing other features such as fold index [47] derived from primary sequence, and solvent accessibility and defined secondary structure from DSSP (Define Secondary Structure of Proteins) profile [48]. We know that under certain conditions of pH, temperature or denaturant concentration, the protein may partially unfold, which may accidentally interact with another unfolding proteins. We speculate that insoluble proteins may be more prone to such behavior by having more segments with negative fold indexes (i.e. likely to be unfold) in their sequences compared to soluble proteins. The DSSP features are used to ensure these segments are on the protein surface and to provide which secondary structures are more likely to be involved. There are a number of solubility prediction schemes based on amyloid-forming propensity of a sequence [30], with β-structure formation underlying this. Whilst such predictors are not always particular effective [31], figure 3.10 supports a view that there are some differences between structural classes, and currently the α/β class performs worst in the 2D non-charged patch analysis. It may be that a guided introduction of β-forming propensity will improve solubility prediction model reported in
here. Other features can also be examined. Finally, the third area to consider, is how protein solubility measured for expression in *E. coli*, matches to expression in other systems, and also how any relatively short time scale expression process relates to the problem of designing for long term storage, and prevention of aggregation, in biopharmaceutical processing. Recent advances in protein stabilisation technology (Arecon, Arestat™) has shown how protein drug (derived from recombinant protein) could have an improved stability and longer shelf-life by formulation approach. Traditional formulation focuses on controlling basic factors such as pH, buffer, stabiliser and antioxidant, while superior stability can be achieved by targeting the degradation pathways in proteins. Such pathways may involve proton exchange, hydrolytic processes, ionic interactions and hydrophobic interactions, the introduction of excipients (inactive substances used as a carrier for the active drug) can prevent or slow down these degradation pathways. Consider a protein drug with exposed hydrophobic residues such as phenylalanine (sidechain with benzene ring) on its surface, the increase of ionic strength of the solution will enhance the hydrophobic interactions between these molecules. An approach to mask this hydrophobic surface with benzene ring molecules with a charge avoid such interaction. The introduction of charges on hydrophobic surface could prevent aggregation, therefore we speculated that similar effect could be also be achieved by adding charge residues on the protein surface by protein engineering to increase solubility, i.e. prevent hydrophobic interactions. In terms of protein expression in eukaryotic systems, the secretory pathway adds back in consideration of subcellular pH variation, as well as, the presence/absence of chaperones. For longer-term storage and formulation, it is likely that partial unfolding, and associated aggregation, is a limiting factor. Some of the ideas presented in here will be relevant for introducing these aspects into solubility/aggregation models.
3.5 References


CHAPTER 4 - Refining the Discrimination of Enzyme and Active Site Identification through Novel Charge-based Calculations

PAPER 3

Pedro Chan¹ (PC), Jim Warwicker¹* (JW)

¹Faculty of Life Sciences, University of Manchester, Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester, M1 7DN, UK

*Corresponding author

JW and PC together designed the study and wrote code for the calculations. The dataset was assembled by PC, and PC performed the calculations and most of the data analysis. PC wrote the manuscript. JW revised the manuscript.

(Article in preparation)
Abstract

A pressing challenge in biology is predicting function of proteins. A general approach to this problem is to assign function to a new protein by matching to homologous sequences of known activity. However, this could become problematic when no sequence homologues are found. Alternative studies using geometric and electrostatics approaches have shown promising results in discriminating enzymes from other protein types, mostly based on the larger cleft size in enzyme. These methods can fail for non-enzyme proteins with large clefts, as is often the case for carbohydrate binding (CB) proteins and periplasmic binding proteins (PBP). Here, we employ electrostatics with two novel charge weighting schemes on a set of 54 CB enzymes, 37 CB non-enzymes and 36 PBPs. We found that the ‘ionisable same sign’ weighting scheme (ISS; uniform charging for all charged atoms of the ionisable groups over protein volume) and ‘pK buried’ weighting scheme (PKB; uniform charging for all charged atoms of the buried ionisable groups over protein volume) can reduce PBPs being falsely predicted as CB enzymes to 20-30% compared to 50% of falsely predicted CB enzymes using the ‘uniform charge’ weight scheme (UNI; uniform charging for all non-hydrogen atoms over protein volume). Further study of 155 enzymes from the Catalytic Site Atlas (CSA) using the ‘standard real charge’ weight scheme (STD; real positive and negative charging and dipoles over protein volume) with grid step of 1.0 Å gave a prediction accuracy close to UNI. Although our current results do not show a better enzyme/non-enzyme separation, it opens up the possibility in recognising strain within the active site (AS) environment, which could be used to categorise the strain pattern for the transition state and could be used to characterise enzymes. A detailed account is given in the discussion section.
4.1 Background

Structural genomics projects aim to expand our structural knowledge of biological macromolecules, especially proteins while lowering the average costs of structure determination [1]. An increase number of protein structures with unassigned function is driving investigation of protein structure/function relationship [2] and the identification of functional sites. The term ‘function’ can be in different aspects such as a molecular function, a cellular role, and/or be part of a functional complex or pathway. Therefore, predicting function from structure alone could be a difficult task if precise details are being questioned, for example, what kind of enzyme, what moieties does it bind, what type of reactions are involved in their functional residues. The general method to assign function to a new protein is by sequence homology but this could fail if no existing homologues are found or the percentage of similarity is extremely low. A more accurate functional inference is using sequence patterns/motifs since proteins that share a common function usually share similarity in part of their sequences/structures and only a few residues are contributing to binding or catalysing within their functional sites. Tools available from these sequence motif databases such as PROSITE [3], PRINTS [4] and BLOCKS [5] can detect useful information about the function of the query protein if matches were found. Structural similarities can infer function of an unknown protein by assigning its secondary structures (or predicted secondary structures if only sequence is given) to fold classes since fold classes are often associated with a particular set of protein functions [6]. But a specific function may not be associated with a fold if the proteins have undergone convergent evolution where enzymes with the same enzyme classification numbers can have entirely different sequences or structural folds [7]. For example, the TIM barrel fold is known to be involved in at least 18 different enzymatic processes [8], thus making this kind of method non-trivial and driving the development of non-alignment-based approaches.

Over the years, alternative methods and tools have been developed to complement sequence and sequence patterns based methods. Several studies have shown that features derived from protein sequences and structures are informative and proteins with similar functions share same features in most cases, which makes function prediction possible without alignment. Sequence-derived features such as secondary structure content, glycosylation site and the number of negatively charged residue have shown to be useful in predicting archaeal enzymes [9].
Structure-derived features such as surface area and density of alpha-carbon atoms have shown similarity in unrelated proteases and when these were trained in a neural network, it gave over 86% accuracy in predicting proteases and non-proteases [10]. Residue conservation derived from multiple sequence alignment (MSA) [11, 12] is able to detect functional important residues in proteins. By applying methods of detecting evolutionary constraints [13], one can identify whether these residues are structural or functional conserved. It has been reported that ligand binding sites tend to lie within the largest clefts of single-chain enzymes in 83% of cases in a study of 67 enzymes [14], which suggested that enzyme active site (AS) can be identified using protein geometry alone. Prediction of enzyme AS based on geometry [15, 16] and electrostatics [17, 18] were developed. The incorporation of residue conservation with cleft detection method has shown a high accuracy in predicting AS residues in enzymes and functional residues in non-enzymes [19, 20]. When 57 of these sequence- or structure-derived features were trained in machine learning, a 77% accuracy is obtained in predicting enzymes and non-enzymes, which shows that the most useful features were secondary structure content, amino acid frequencies, number of disulphide bonds and size of the largest cleft [21].

Our previous study [22] in distinguishing enzymes and non-enzymes has shown that using continuum electrostatics with UNI alone can give sensitivity:specificity ratio of 80%:80% in a set of 156 enzymes and 174 non-enzymes. This method assigns positive charges to all heavy atoms (except hydrogen atoms) of the protein and apply Finite Difference Poisson-Boltzmann (FDPB) to calculate electric potentials (measured in millivolts (mV)) on the surface of the protein. The point on the protein which has the highest electric potential is termed the ‘peak potential’ (POTL) and is recognised as the largest cleft. This is an alternative way to measure the cleft size of the protein and the threshold of 780 mV gave the best result for separating enzymes from non-enzymes. However, non-enzymes with high POTL (>780 mV) were predicted as enzymes and most of these were CB proteins. Therefore there is a need to refine our current method.

Experimental findings and theoretical considerations suggested that residues involved in catalytic reaction often under energetically unfavourable environments [23]. These are usually ionisable residues and their pKas may be differ from their native solvent state. Following this observation, we investigate whether focusing on POTL produced by ionisable groups could
improve enzyme/non-enzyme discrimination. Our first study compared the CB enzymes/CB non-enzymes discrimination results obtained from UNI, ISS and PKB. ISS and PKB did not show better separation than the UNI but when POTL from all three schemes were multiplied together (a way to bring these properties into a higher dimension for analysis), a slight improvement in separating CB non-enzyme is obtained. A set of 37 PBPs is included for testing ISS and PKB since PBPs are also non-catalytic, similar to CB non-enzymes. PBPs bind to a variety of small solutes such as sugars, metal ions and amino acids and many of them act as receptors or transporters. PKB outperformed UNI in discriminating 20% more of PBPs in this test.

The second study investigates whether STD could improve enzyme AS prediction with new electrostatic features and new strain measuring schemes in a set of 155 enzyme structures. In previous study [17], although UNI outperformed STD in AS prediction with a high accuracy, it doesn't allow us to look further into charges within the AS that may be useful in categorising electrostatic strain patterns. This analysis compare prediction results generated from UNI and ISS as well as from different electrostatic features such as POTL and ‘peak electric field’ (EFLD) generated from STD. Prediction results were collected in two steps: (1) electrostatic features (and its location/coordinates) were given by charges and dipoles (as in STD) or pseudo uniform charges (as in UNI and ISS) from the enzyme structure and (2) enzyme AS prediction was based on the location to the centroid (the average location of the true AS residues of the enzyme obtained from CSA). The enzyme set was adopted from previous publication [17]. Amongst all electrostatic features from STD, P2-POTL gave an encouraging result when comparing to UNI and ISS. Within 15.0 Å radius of the centroid, STD made correct prediction for 123 enzymes whereas 127 and 131 were predicted in UNI and ISS respectively. Although STD is still outperformed by other schemes, we have gained an improvement by using (1) a smaller grid step and (2) a patch method, in combination to sample electrostatic strains at a closer distance to the molecular surface in a protein cleft. Future improvement of our approach will allow us to investigate the possibility of predicting enzymes by charge and dipole patterns from their AS.
4.2 Methods

4.2.1 Datasets

4.2.1.1 CB Enzymes and CB Non-enzymes
Protein Data Bank (PDB) [24] was searched on 02 Feb 2006 for X-ray structures which gave 27616 protein entries. Carbohydrate-protein complexes containing any of these six sugar moieties: glucose (GLC), galactose (GAL), mannose (MAN), fucose (FUC), N-acetylglucosamine (NAG) and sialic acid (SIA) were searched from the protein list which gave 1825 PDB IDs. This choice of the six sugar moieties follows previous work [25]. Of these, 715 glycosylated proteins were discarded because these carbohydrates were covalently bound to the proteins and were not in the protein’s functional site. Glycosylation site of the protein and the attached molecules were indicated in the MODRES and LINK line of the PDB file respectively. The rest of 1110 proteins were examined for the distance from their ligands. The distance is calculated between oxygen atoms of each ligand and nitrogen atoms of its protein. A contact cut-off distance of 3.0 Å was used (slightly less than the longest possible hydrogen bond length of 3.44 Å), and any ligand with at least one hydrogen bond to the protein was considered as protein bound. The resultant consists of 1063 ligand bound proteins. They were compared to a list of 16715 enzymes with known enzyme classification (EC) numbers obtained from the integrated database retrieval system, DBGET/LinkDB [26]. This gave 629 CB enzymes and 434 CB non-enzymes. Single chains were used in the analysis to avoid problems in distinguishing equivalent peaks and interchain binding site in multimeric proteins, therefore the CB enzyme and CB non-enzyme sets were culled by chains using PISCES [27], a sequence culling server, to eliminate sequences that were more than 25% similar. (In chapter 2 and 3, the sequence similarity was set at 90% since these studies required as many proteins as possible but not identical chains, which allows residue variation on a common fold. However, in the current study, we required unique protein chains to test with our predicting schemes.) This gave 139 and 115 single chains in the CB enzyme and CB non-enzyme sets respectively. The protein description and publication abstract were manually checked for CB non-enzymes for misclassified proteins. A total of 163 protein chains were removed from both sets. For example: 1CLE, a cholesterol esterase is not complexed with a ligand. Finally, there were 54 and 37 protein chains in the CB enzyme and CB non-enzyme set respectively. The protein names, PDB
IDs and their closely bound ligands can be found in the Appendices. (Please note: During this study, certain PDB files got modified and new ligand names were introduced into these files, which caused the absence of the bound ligands for 10 proteins. For example: GAL was replaced by GLA in 1pie, thus bound ligands were missing for these proteins. Ligands in these proteins were manually identified and a separate run of the program was given to ensure these ligands were protein bound. These PDB IDs are marked with an asterisk and a separate table is included in the Appendices for full explanation.) Figure 4.1 gives a schematic view of the CB datasets selection. Additional File 4-1 contains the dataset and results for this CB proteins study.

![Figure 4.1: A schematic diagram of the data selection for CB enzyme and CB non-enzyme sets](image-url)
4.2.1.2 Periplasmic Binding Proteins

A list of 335 PDB identifiers (IDs) of PBPs was obtained from the two folds categories: Periplasmic binding protein-like I and II from the Structural Classification of Proteins (SCOP) database [28]. The PBP list was searched against 38980 PDB IDs for ligand binding structures. Of these, 268 ligand bound PBPs were retained. The rest of 65 ligand free PBPs and 2 NMR structures were discarded. The 268 PBP structures (consist of 408 protein chains) were sent to PISCES for sequence culling at 25% cut-off, which gave 36 chains for the PBP set. Although the cut-off percentage is low, this does not represent a complete non-redundancy since the general structural fold is conserved in the two PBP classes [29] but their sequence diversity within their superfamily is high.

4.2.1.3 Enzymes

A list of 155 enzymes from the reported publication [17] was used. In brief, 178 enzymes were selected from the CSA database [30], which is a set of non-homologous enzymes of known structure with a well-defined AS [31]. The set was reduced to 159 upon removal of homologues [32]. Four proteins (1a6f, 1b66, 1dqa, 1bzy) were further excluded. The first one being modulates rather than defines catalysis; the active sites of the next two were formed by more than one polypeptide [17]; and the last one was a duplicate. The current study uses single chains to avoid problems in distinguishing equivalent peaks. Additional File 4-2 contains the dataset used in the enzyme study.

4.2.2 Calculations

4.2.2.1 Charge Weighting Schemes

Four types of charge weighting schemes were used in our studies. STD assigns a +1 charge to positive charged atoms (Arg and Lys) and a -1 charge to negative charged atoms (Asp and Glu). UNI assigns positive charges to all heavy atoms (except hydrogen) of the protein such as carbon, nitrogen and oxygen. ISS assigns positive charges to the charged atoms of the functional group of all ionisable residues of the protein. PKB assigns positive charges to the charged atoms of the functional group of ionisable residues that have large pKa shifts within the
protein. Figure 4.2 gives a simplified graphical representation of these charge weighting schemes.

Figure 4.2: Schematic view of the charge weighting schemes
(a) STD; (b) UNI; (c) ISS; and (d) PKB.

4.2.2.2 Electrostatic Calculations
The current work employed the FDPB method to calculate strain from clefts of the protein structures. The electrostatic strain can be represented by electrostatic features such as electric potential and electric field. The PDB file of a protein structure is served as the input of our program. At first, the molecular surface (also known as the solvent accessible surface) of the protein is recognised by rolling a water probe with radius of 1.4 Å. It is defined as the contact surface formed between the van der Waals envelope of the protein atoms and the centre of the probe. Depending on the choice of weighting scheme, a subset of protein atoms were assigned with positive charges for UNI, ISS and PKB or a mix of positive and negative charges for STD (-1 assigned to Asp, Glu and C-terminus, +1 assigned to Arg, Lys and N-terminus, +0.5
assigned to His). The protein is then placed into a three dimensional grid with grid spacing = 2.0 Å, \(\varepsilon_s\) (Solvent dielectric) = 78.4, \(\varepsilon_p\) (Protein dielectric) = 4.0 and ionic strength = 0.15 Molar or as indicated in the result table. Electrostatic terms such as charge density and dielectric constant of each charged atom of the protein within a grid cube is assigned to the central grid point. The electric potential and electric field of these grid points were calculated by FDBP on the second nearest neighbour shell (2NN) of the molecular surface (solvent side), taking nearby interactions from the six nearest grid points into account. The first nearest neighbour was not used to avoid irregularities at the protein/solvent boundary itself. Electric potentials of the protein were ranked and the one with the highest mV was termed peak potential (POTL), which was used to represent the largest cleft. The same was done for the electric field.

4.2.2.3 Enzyme AS Prediction

AS centre for each entry of the 155 enzyme set is the location averaged over C\(\beta\) atom positions of the listed catalytic residues from the CSA [32] and is termed the centroid. A protein is predicted as enzyme if the position of the POTL is within a 10 Å radius of the centroid. A radius of 15 Å was also used. Apart from the POTL and EFLD, we also represent the strain of a cleft by type 1 patch peak potential (P1-POTL), type 2 patch peak potential (P2-POTL) and patch peak electric field (P-EFLD). POTL and EFLD collect strain from a single point but the strain collected from P1-POTL, P2-POTL or P-EFLD was averaged over a number of points from the ‘patch probe’ (Table 4.1 shows the number of points in each ‘patch probe’). P1-POTL and P2-POTL differ in their number of points being used and the number of points within the ‘patch probe’ is relative to its size. Briefly, the procedure employs a sphere with 10 Å radius (also tested with 12 and 14 Å) to detect solvent holes by rolling over the protein surface. Once a cleft is detected, the points on the ‘patch probe’ are used to record the electric potential or electric field from the 2NN, which is 4.0 Å from the molecular surface (if a 2.0 Å grid is used). A patch potential is an average of several potentials within the ‘patch probe’ (radius of 12 Å performed best in our case). Figure 4.3 gives a graphical representation.
<table>
<thead>
<tr>
<th>Patch Probe Radius (Å)</th>
<th>P1-POTL (Points)</th>
<th>P2-POTL (Points)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>60</td>
<td>157</td>
</tr>
<tr>
<td>12</td>
<td>72</td>
<td>226</td>
</tr>
<tr>
<td>14</td>
<td>84</td>
<td>307</td>
</tr>
</tbody>
</table>

Table 4.1: Number of points presented in each ‘patch probe’ (10, 12 and 14 Å) used for collecting patch potentials.

Figure 4.3: 2D representation of the patch method (not to scale)
4.3 Results and Discussion

4.3.1 Enzyme/Non-enzyme Discrimination

A set of 91 protein structures containing 54 CB enzymes and 37 CB non-enzymes was served as an initial test for our enzyme/non-enzyme discrimination problem. Three different weighting schemes were used to calculate the POTL for the protein structures and the resultant values from each weighting scheme were analysed by their cumulative frequency. A threshold was placed at the point where the largest separation was shown between the enzyme and non-enzyme sets. The resultant data from the plot was then divided into four groups: (1) true positives (TP) for correctly predicted CB enzymes (POTL > threshold); (2) false positives (FP) for CB non-enzymes predicted as CB enzymes (POTL > threshold); (3) true negatives (TN) for correctly predicted CB non-enzymes (POTL < threshold); (4) false negatives (FN) for enzymes predicted as non-enzymes (POTL < threshold). Thus, sensitivity (TP / (TP + FN) ) and specificity (FP / (FP + TN) ) for each weighting scheme could be deduced and compared. A comparison was later made with a set of PBPs.

4.3.1.1 UNI Compares with Previous Work

Our first result of discriminating CB enzymes and CB non-enzymes using UNI showed a similar trend to our previous work [22] of enzyme/non-enzyme (ENZ/NON-ENZ) discrimination. Each datum on the plot (Figure 4.4) represents the cumulative percentage of the number of proteins equal or below the peak potential value. Data of CB enzyme and non-enzyme sets were slightly to the right, indicating they have a lower percentage at each peak potential. The threshold to divide them is set at 840 mV where the largest difference is shown in figure 4.5. In the CB enzyme set, 90% were predicted correctly and the remaining was classified as CB non-enzymes due to their small POTLs, which are relative to their short chain length of less than 200 AA residues. 24% of proteins in the CB non-enzyme set were predicted as CB enzymes mainly due to their large cleft sizes. It is this feature that causes the false positives problem in the previous ENZ/NON-ENZ problem.
Figure 4.4: Comparing ENZ/NON-ENZ and CB enzyme/CB non-enzyme results in UNI
ENZ in dark blue; NON-ENZ in light purple; CB enzyme in yellow and CB non-enzyme in light blue.

Figure 4.5: CB enzyme/CB non-enzyme discrimination in UNI
Threshold at 840 mV
Largest difference 68%
4.3.1.2 ISS and PKB
Cumulative plots for ISS and PKB (not shown) showed similar pattern as UNI plot. The thresholds, TP, FN, TN, FP and Sensitivity:Specificity ratio of the three schemes are shown in table 4.2. The ROC plot (Sensitivity : 1-Specificity) is shown in figure 6. The UNI scheme is best for such discrimination since it separates more CB enzymes and CB non-enzymes. The ISS scheme is less effective for discriminating CB enzymes and the PKB scheme is less effective for discriminating CB non-enzymes.

<table>
<thead>
<tr>
<th>Weighting Schemes</th>
<th>POTL Threshold</th>
<th>TP</th>
<th>FN</th>
<th>TN</th>
<th>FP</th>
<th>Sensitivity:Specificity ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNI</td>
<td>840</td>
<td>49</td>
<td>5</td>
<td>28</td>
<td>9</td>
<td>0.91:0.76</td>
</tr>
<tr>
<td>ISS</td>
<td>240</td>
<td>46</td>
<td>8</td>
<td>28</td>
<td>9</td>
<td>0.85:0.76</td>
</tr>
<tr>
<td>PKB</td>
<td>37</td>
<td>49</td>
<td>5</td>
<td>22</td>
<td>15</td>
<td>0.91:0.60</td>
</tr>
</tbody>
</table>

Table 4.2: Result summary for the three weighting schemes for 54 CB enzymes and 37 CB non-enzymes.

Figure 4.6: ROC plot for CB enzyme/CB non-enzyme discrimination in three weighting schemes
4.3.1.3 Combined Peak Potentials to Discriminate CB enzyme/CB Non-enzyme

The ISS and PKB schemes did not show significantly better performance on CB enzyme/CB non-enzyme discrimination than the UNI scheme. Next we looked at whether the combined peak potentials (C-POTL; multiplying the peak potential from different weighting schemes for each protein) could improve their separation. Their results and performance are summarised in table 4.3. The various C-POTLs did not perform substantially better overall than using individual weighting scheme. But they all (apart from UNI x ISS) showed an increase in their specificity although with a drop in their sensitivity. This provide a slightly better threshold for identifying CB non-enzymes, so that the false positive prediction rate (1-Specificity), for our original observation for CB non-enzymes is improved. Parameters for the UNI scheme was set to give the best performance as previously described [22]. Since the method assigned positive charges to all heavy atoms of the protein is for mimicking a geometrical environment, therefore, we did not attempt to change any parameters such as dielectric constant or ionic strength because they will not affect the results. Only less positive charges are assigned to proteins when using ISS and PKB schemes.

<table>
<thead>
<tr>
<th>Combined Weighting Schemes</th>
<th>C-POTL Threshold (mV^2)</th>
<th>TP</th>
<th>FN</th>
<th>TN</th>
<th>FP</th>
<th>Sensitivity: Specificity ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNI x ISS</td>
<td>180000</td>
<td>48</td>
<td>6</td>
<td>27</td>
<td>10</td>
<td>0.89:0.73</td>
</tr>
<tr>
<td>UNI x PKB</td>
<td>34000</td>
<td>46</td>
<td>8</td>
<td>29</td>
<td>8</td>
<td>0.85:0.78</td>
</tr>
<tr>
<td>ISS x PKB</td>
<td>14000</td>
<td>40</td>
<td>14</td>
<td>33</td>
<td>4</td>
<td>0.74:0.89</td>
</tr>
<tr>
<td>UNI x ISS x PKB</td>
<td>9300000 (mV^3)</td>
<td>44</td>
<td>10</td>
<td>31</td>
<td>6</td>
<td>0.81:0.84</td>
</tr>
</tbody>
</table>

*Table 4.3: Result summary for C-POTLs for 54 CB enzymes and 37 CB non-enzymes.*
4.3.1.4 Combined Peak Potentials to Discriminate ENZ/NON-ENZ sets

Since the C-POTLs have shown slight improvement in predicting CB non-enzymes, we applied the same approaches to RGJW ENZ and NON-ENZ set. Table 4.4 summarised the results. Neither sensitivity nor specificity from C-POTLs showed improvement when compared to the UNI scheme result.

<table>
<thead>
<tr>
<th>Combined Weighting Schemes</th>
<th>C-POTL Threshold</th>
<th>TP</th>
<th>FN</th>
<th>TN</th>
<th>FP</th>
<th>Sensitivity: Specificity ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNI</td>
<td>780 (mV)</td>
<td>124</td>
<td>29</td>
<td>140</td>
<td>34</td>
<td>0.81:0.80</td>
</tr>
<tr>
<td>UNI x ISS</td>
<td>170000 (mV²)</td>
<td>123</td>
<td>30</td>
<td>138</td>
<td>36</td>
<td>0.80:0.79</td>
</tr>
<tr>
<td>UNI x PKB</td>
<td>36000 (mV²)</td>
<td>107</td>
<td>46</td>
<td>127</td>
<td>47</td>
<td>0.70:0.73</td>
</tr>
<tr>
<td>ISS x PKB</td>
<td>9000 (mV²)</td>
<td>111</td>
<td>42</td>
<td>112</td>
<td>62</td>
<td>0.73:0.64</td>
</tr>
<tr>
<td>UNI x ISS x PKB</td>
<td>9000000 (mV³)</td>
<td>109</td>
<td>44</td>
<td>135</td>
<td>39</td>
<td>0.71:0.76</td>
</tr>
</tbody>
</table>

*Table 4.4: Result summary for C-POTLs for RGJW enzymes and non-enzymes set.*

4.3.1.5 PBPs

To further investigate the potential to use ISS and PKB in ENZ/NON-ENZ separation, a study was made with PBPs. These proteins are usually found in the periplasmic space in the Gram-negative bacteria and are responsible for binding and transporting a variety of sugars, amino acids and nutrients [33]. Also, these proteins are not catalytic and would be a suitable as another test set of non-enzyme binding proteins. Figure 4.7 shows a cumulative frequency plot for CB enzymes, CB non-enzymes and PBPs. The PBP data sits between CB enzymes and CB non-enzymes, indicating they have a higher proportion of proteins with small cleft size than CB enzymes but not as many as CB non-enzymes. Their result summary is shown in table 4.5. The C-POTL method has shown improvement by reducing the number of PBPs falsely predicted as CB enzymes when compared to UNI (Table 4.4). The ISS and ISS x PKB schemes were the best two with an accuracy \(((TP+TN)/(TP+FN+TN+FP))\) of 0.73 and 0.69 respectively.
Figure 4.7: Cumulative plot for the discrimination between CB enzymes, CB non-enzymes and PBPs in UNI

<table>
<thead>
<tr>
<th>Weighting Schemes</th>
<th>POTL Threshold</th>
<th>TP</th>
<th>FN</th>
<th>TN</th>
<th>FP</th>
<th>Sensitivity: Specificity ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNI</td>
<td>960</td>
<td>44</td>
<td>10</td>
<td>18</td>
<td>18</td>
<td>0.81:0.50</td>
</tr>
<tr>
<td>ISS</td>
<td>320</td>
<td>38</td>
<td>16</td>
<td>28</td>
<td>8</td>
<td>0.70:0.78</td>
</tr>
<tr>
<td>PKB</td>
<td>45</td>
<td>41</td>
<td>13</td>
<td>24</td>
<td>12</td>
<td>0.76:0.67</td>
</tr>
<tr>
<td>UNI x ISS</td>
<td>290000</td>
<td>41</td>
<td>13</td>
<td>22</td>
<td>14</td>
<td>0.76:0.61</td>
</tr>
<tr>
<td>UNI x PKB</td>
<td>430000</td>
<td>42</td>
<td>12</td>
<td>21</td>
<td>15</td>
<td>0.78:0.58</td>
</tr>
<tr>
<td>ISS x PKB</td>
<td>14500</td>
<td>40</td>
<td>14</td>
<td>26</td>
<td>10</td>
<td>0.74:0.72</td>
</tr>
<tr>
<td>UNI x ISS x PKB</td>
<td>18300000</td>
<td>35</td>
<td>19</td>
<td>28</td>
<td>8</td>
<td>0.65:0.78</td>
</tr>
</tbody>
</table>

Table 4.5: Result summary for peak potentials for 54 CB enzymes and 36 PBP non-enzymes.
4.3.2 Enzyme AS Prediction

An enzyme set of 155 was tested with the STD scheme using different sets of parameters including grid step, patch size, pH and protein/solvent dielectric and ionic strength in order to benchmark with UNI. An additional test using ISS was later included. Table 4.6 lists the best prediction results using grid step of 1.5 and 2.0 Å. Table 4.7 lists the best prediction results using grid step of 1.0 and 2.0 Å.

<table>
<thead>
<tr>
<th>Weighting Schemes</th>
<th>Grid step (Å)</th>
<th>pH</th>
<th>Dielectric (protein, solvent)</th>
<th>POTL 10 Å</th>
<th>POTL 15 Å</th>
<th>P1-POTL 10 Å</th>
<th>P1-POTL 15 Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNI</td>
<td>2</td>
<td>7</td>
<td>80, 80</td>
<td>99</td>
<td>128</td>
<td>25</td>
<td>80</td>
</tr>
<tr>
<td>UNI</td>
<td>1.5</td>
<td>7</td>
<td>80, 80</td>
<td>97</td>
<td>129</td>
<td>41</td>
<td>103</td>
</tr>
<tr>
<td>STD</td>
<td>2</td>
<td>5.5</td>
<td>1, 80</td>
<td>85</td>
<td>109</td>
<td>30</td>
<td>71</td>
</tr>
<tr>
<td>STD*</td>
<td>1.5</td>
<td>5.5</td>
<td>1, 80</td>
<td>75</td>
<td>105</td>
<td>44</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 4.6: Parameters and results for predicting AS from the 155 enzyme set using grid step of 1.5 and 2.0 Å.

Probe radius of 1.4 Å, patch size of 10 Å and ionic strength of 0.15 Molar were used in the calculation shown in the table. a 0.01 Molar was used instead.

<table>
<thead>
<tr>
<th>Weighting Schemes</th>
<th>Grid step</th>
<th>pH</th>
<th>Dielectric (protein, solvent)</th>
<th>POTL 10 Å</th>
<th>POTL 15 Å</th>
<th>P2-POTL 10 Å</th>
<th>P2-POTL 15 Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNI</td>
<td>2</td>
<td>5.5</td>
<td>1, 80</td>
<td>99</td>
<td>128</td>
<td>103</td>
<td>127</td>
</tr>
<tr>
<td>UNI</td>
<td>1</td>
<td>5.5</td>
<td>1, 80</td>
<td>34</td>
<td>89</td>
<td>80</td>
<td>126</td>
</tr>
<tr>
<td>STD</td>
<td>2</td>
<td>5.5</td>
<td>1, 80</td>
<td>77</td>
<td>100</td>
<td>63</td>
<td>84</td>
</tr>
<tr>
<td>STD</td>
<td>1</td>
<td>5.5</td>
<td>1, 80</td>
<td>50</td>
<td>98</td>
<td>87</td>
<td>123</td>
</tr>
<tr>
<td>ISS</td>
<td>2</td>
<td>5.5</td>
<td>1, 80</td>
<td>34</td>
<td>100</td>
<td>90</td>
<td>111</td>
</tr>
<tr>
<td>ISS</td>
<td>1</td>
<td>5.5</td>
<td>4, 80</td>
<td>67</td>
<td>110</td>
<td>95</td>
<td>131</td>
</tr>
</tbody>
</table>

Table 4.7: Parameters and results for predicting AS from the 155 enzyme set using grid step of 1.0 and 2.0 Å.

Probe radius of 1 Å, patch size of 12 Å, pH of 5.5 and ionic strength of 0.15 Molar were used in the calculation shown in the table.
4.3.2.1 Compare STD to UNI and ISS

Single peak potential method was best when using UNI, which correctly predicted 99 and 128 enzymes to within 10 and 15 Å to the centroid respectively (See table 4.6 and 4.7). The number of predicted enzyme AS dropped when the grid step and other parameters were reduced. This is true in all POTL results in both UNI and STD in general. The opposite trend was observed in the P1-POTL results with grid step of 1.5. This is due to the smaller grid step being used, which reduced the distance (from 4.0 to 3.0 Å from the molecular surface) to collect electric potential. This shows that UNI is best performed when the potential is sampled at 4.0 Å from the molecular surface, while P1-POTL improves when the electric potentials are sampled from a closer distance. The same trend is observed when the grid step is adjusted to 1.0 Å with patch type, P2-POTL (Table 4.7). It has successfully increase the enzyme number (within 15 Å) to 123 in STD, which matched closely to the UNI scheme. The possible reason for such results is that for single peak potential method, as grid step reduces, it has a higher chance to pick up signal from small solvent holes and not at the AS, thus performance drops. But for patch potential, the final value is averaged over many neighbouring values, therefore less susceptible to small holes. Further tests were made with grid step of 0.5 Å but no further improvement was found. ISS was later added to the test and surprisingly it topped the chart with 131 enzymes in the P2-POTL (within 15 Å). From the results shown in table 4.6 and 4.7, we found that the ISS performed slightly better than UNI in the P2-POTL (within 15 Å). For each enzyme, DIST, which defines as the distance between the predicted AS location and the averaged location of the centroid is recorded for each scheme (See Appendices). Their scatter plots were shown in figure 4.8, in order to show their correlations but only a low correlation was found in each plot.
Figure 4.8: The DIST of the enzyme set between different weighting schemes in XY scatter plots

Data fall below the red horizontal line and to the left of the green vertical line are within the 15 Å threshold. Weighting schemes with the highest predicted enzyme AS number were plotted against each other. (a) UNI (POTL) vs STD (P2-POTL). (b) UNI (POTL) vs ISS (P2-POTL).
4.3.2.2 Failures for AS Location

After our manual examination of the result data, we found that 99 out of 155 enzymes with their AS predicted correctly (within 15 Å) using any of the three schemes. The rest of 56 enzymes that fall outside the 15 Å radius (in any of these schemes) are listed in table 4.8. Five of these failed in all schemes, for example: 1gpm, a GMP synthetase that has a bilobal active site, which makes the averaged AS far from the true location (Figure 4.9(a)); 1pkn, a muscle pyruvate kinase that has its AS located on the interface between two domains (Figure 4.9(b)). Figure 4.10 shows a schematic view of how these schemes could fail if the AS is being located in a shallow cleft or near the protein surface using 1ab8 as an example.

<table>
<thead>
<tr>
<th>Failure in</th>
<th>PDB IDs and comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNI, STD, ISS</td>
<td>1ab8 (AS on protein surface), 1kfu (formerly 1dkv; AS on protein surface), 1gpm (Bilobal active site), 1gpr (AS near surface of protein), 1pkn (not quite a cleft, AS locates on interface between two domains)</td>
</tr>
<tr>
<td>UNI, STD</td>
<td>1alk, 1mek, 4kbp, 1fua 1dco (AS on protein surface)</td>
</tr>
<tr>
<td>UNI, ISS</td>
<td>1cqq, 1dpn, 1hxq, 1lxa, 1mhy, 1nzy, 1uvox, 1ytw 1d4a (AS on protein surface)</td>
</tr>
<tr>
<td>STD, ISS</td>
<td>1fro, 1gog, 2pda, 3eca 2adm (formerly 1adm, AS on protein surface)</td>
</tr>
<tr>
<td>UNI only</td>
<td>1b93, 1chd, 1fui, 1gtp, 1i7d, 1jdw, 1kra, 1nba, 1pn1, 3csm, 3pca 2bmi (formerly 1bmi, AS on protein surface)</td>
</tr>
<tr>
<td>STD only</td>
<td>1a26, 1ae7, 1b3r, 1b6b, 1bol, 1bzy, 1cd5, 1dii, 1do8, 1dj, 1ivh, 1ra2, 1req, 2cpo 1hfs (one AS on protein surface, one buried under three HIS residues)</td>
</tr>
<tr>
<td>ISS only</td>
<td>1cb8, 1e2a, 1foh, 1gpa 2acy (AS on protein surface)</td>
</tr>
</tbody>
</table>

Table 4.8: Enzymes with their predicted AS fall outside the 15 Å threshold.
Figure 4.9: Examples of enzymes with their AS that fall outside the 15Å threshold
Proteins are coloured in white and their true AS residues are in various colours.
Twelve enzymes with their predicted AS fall outside the 15 Å threshold in UNI but not in others, for example: 2bmi (formerly 1bmi), a metallo-beta-lactamase with its AS situated on the protein surface, which exposed to the solvent (Figure 4.9(c)). Although its peak potential is 823 mV, further investigation shows that the predicted location is different from the true AS. Fifteen enzymes with their predicted AS fall outside the 15 Å threshold in STD but not in others, for example: 1hfs, a catalytic domain of human fibroblast stromelysin-1 with two AS residues (Figure 4.9(d)). One being located on the protein surface and the other one were buried under the histidine triad (HIS 201, 205 and 211). A zinc ion is reported as the co-factor for the catalytic activity and is held by the histidine triad. Five were failed in the ISS but not in other schemes, for example: 2acy, an acyl-phosphatase has its AS exposed to the surface of the protein (Figure 4.9(e)). The literature suggested that the catalysis is involved with a chloride ion, a sulphate ion and a water molecule.

Figure 4.10: All schemes failed to predict the AS within 15Å of the centroid of 1ab8
ARG 1029 is the true AS. (a) Front view; (b) Rear view
4.4 Conclusions

4.4.1 Distinguishing CB enzymes from CB Non-enzymes and PBPs

The ISS and PKB used in separating CB enzymes from CB non-enzymes did not give better results than the UNI. ISS gave the same specificity of 0.76 as UNI with a reduced sensitivity, while PKB gave the same sensitivity of 0.91 with a reduced specificity. ISS and PKB depend on finding a cleft with the highest peak potential in a protein, which reflect the number of ionisable groups and the buried pKas respectively. Although they are marginally less effective than UNI, they showed that electrostatic strain created by charges are also possible to predict enzymes from non-enzymes. Within the prediction results, we observed that PKB can detect enzymes that were previously predicted as non-enzymes in UNI scheme.

The predictive power for ISS and PKB was more observable when they are used to predict PBPs. Both schemes detect more PBPs than UNI. ISS has a specificity of 0.78, while it is 0.5 for UNI. A possible explanation is that proteins from our PBP set shared similar structural fold (already mentioned in the Methods section) and thus, their cleft sizes are more or less similar to each other. The prediction from UNI is likely to be bias towards enzymes if the majority of PBPs has a large cleft in general. However, charges are less required by binding sites of PBPs compared to enzymes in general, therefore we see a better performance in ISS and PKB.

4.4.2 Predicting Enzymes using Detailed Charge Distribution

The aim for the enzyme set study is to re-visit the more detailed real charge (STD) method to see whether different settings of the parameters would give any better performances in enzyme prediction. The single peak potential method used with UNI predicted AS (within 15 Å) correctly in 129 out of 155 enzymes with a grid step of 2 Å. In order to obtain a stronger signal from the AS, we sampled the peak potential at a closer distance to the molecular surface by lowering the grid step to 1 Å. However, the performance dropped due to the pickup of other signals from small solvent holes. Alternatively, the patch potential method perform better with small grid step. A number of potentials was sampled within a patch and they were averaged to give a final potential. This method worked less effective with UNI but it predicted 123 out of 155 enzyme AS
(within 15Å) correctly using STD. However, the peak potential method with UNI still outperform other settings for STD. If we have obtained a better performance from STD, we would have begun our investigation into the charge template idea. From transition state theory, we learned that there are electrostatic strains within the enzyme AS during the transition state. Since STD is the only scheme that allows us to look into real charges as well as dipoles, the potential maps we collected from our program can be used to characterise these electrostatic strains. We believe that there are charges and dipole patterns lie within the active sites of proteins, that may be able to separate different kinds of enzymes. Therefore, refining our current approach continues to be part of our future goal. For a simple refinement, it would be possible to re-visit the single peak method by implementing a scheme to eliminate signals from small solvent holes, which we believed was the cause for noises in the sampling of peak potential using small grid step of 1.0 Å. And for long term solution, there is a need to step back and investigate a different solution to the problem because it seems that our program has reached its limit for giving better results by varying existing parameters. It is evident that for enzymes where their active sites (AS) are located in a shallow cleft or protein surface would be missed by our program. This is because their potential values would be too low to become top of the potential list. It would be beneficial if this kind of enzymes could be further investigated. For example: all the potentials found in a single protein could be compared to each other with other structural or electrostatic information within a certain radius, which provide a ratio feature (e.g. measuring the number of charges within 15 Å radius) that would be used to represent such surface or cleft. If any of these ratio features are able to highlight the AS, then it could be used as a second filter to identify the correct AS in those enzymes that we have failed on.
4.5 References


CHAPTER 5 - Conserved Residue Colouring Web-tool: Identification of Residues that are Evolutionary Conserved

PAPER 4

Pedro Chan¹ (PC), Salim Bougouffa¹ (SB), Richard Greaves¹ (RG) and Jim Warwicker¹* (JW)

¹Faculty of Life Sciences, University of Manchester, Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester, M1 7DN, UK

*Corresponding author

PC created the web-server application and wrote the manuscript. PC and RG supplied computer code and PC and SB provided some website code. JW directed and design of the application and revised the manuscript.

(Article in preparation)
Abstract

Conserved Residue Colouring (CRC) Web-tool is an easy-to-use online program for predicting evolutionary conserved residues on the single chain of a protein structure that may be of functional importance. The resultant structure is displayed on Jmol, a popular molecular viewer, with the highly conserved residues coloured in red and the least conserved ones in blue. CRC is a two-stage process. It first retrieves sequence homologues for the protein chain and aligns the sequences using PSI-BLAST, followed by the calculation of conservation score for each residue. The alignment and conservation score, part of the CRC has been shown to improve geometry-based enzyme active site prediction [1]. Recently, this geometry-based prediction method has become available as a web-tool [2] with CRC being part of it. Since CRC itself has the potential to identify evolutionary conserved residues based on Multiple Sequence Alignment (MSA), it could be used independently without knowing the cleft location of a protein and not restricted to the active site/functional site problem. Therefore, it is now available as a stand-alone tool for broader usage.

CRC is accessible via these links:

http://personalpages.manchester.ac.uk/staff/j.warwicker/resources.html

http://www.bioinf.manchester.ac.uk/~mqbpgpc4/
5.1 Background

The number of protein structures have grown rapidly since the launch of structural genomics (SG) projects. SG projects aim to determine all unique and non-redundant protein structures from human and from a selection of model organisms including C. elegans and M. tuberculosis for clinical studies and to expand our knowledge in proteins [3, 4]. As of 11 May 2010, the SG centres have deposited 9051 structures to the Protein Data Bank (PDB) [5], representing 13.9% of the whole database. Among these proteins, there are currently around 2000 SG structures which have unknown functions. Efforts have been put into the development of algorithms/programs to annotate these proteins at residue level and infer function/functional sites using information from related proteins [6-8]. A general approach to detect a binding site in protein is by using sequence information via MSA [9-13]. However, these approaches may fail if the protein of interest shares no similarity to any currently known protein sequences/structures [14].

Alternative methods have been developed to find functional sites in the absence of residue conservation. These methods range from geometry features such as clefts [15], energetic properties [16] to proximity of residues to the protein centre [17] and structural template matching for active site [18]. Studies that use a combination of geometrical, evolutionary such as phylogenetic information and stability-related information have also been proposed [19, 20].

Previous work by this group [1] used electrostatic peak potential to identify functional site near the surface of a protein. However, the best prediction was the use of a uniform charge weighting scheme, which assign positive charges to all atoms except hydrogen, effectively mimicking a geometry method for detecting cleft size. With the incorporation of the CRC program, which calculates residue conservation using MSA, the accuracy was further improved. Recently, this work has become available as a web-tool [2].

CRC calculates residue conservation for a protein from MSA by using the C-trident formula, which accounts for three important variables (symbol diversity, stereochemical diversity and gaps) for each residue of the protein (see the implementation section for further explanation) [21], and colours the protein structure according to the residue conservation. Alternative ways to estimate residue conservation have been proposed such as the use of position-specific scoring.
matrix (PSSM) profile or entropy of residues [22]. Functional site prediction tools are also found to combine residue conservation with other features such as ligand-specified residues [23], secondary structure and solvent accessibility [24], phylogenetic tree information [25] and protein cavities conservation [26]. Other tools are also available but are limited to sequence input in FASTA format [27] or the input is limited to pre-existing PDB files only [28].

So far, the most popular web-tool is Consurf [25]. Like CRC, it maps residue conservation onto the surface of a protein structure by allowing user to choose different algorithms and parameters in part of the workflow. For example, the user can manually select protein sequences from the PSI-BLAST output for building the MSA as well as choosing different programs such as MUSCLE [29] or CLUSTALW [30] to perform MSA. Over the years, Consurf has enhanced its performance by the implementation of new approaches such as the replacement of “Evolutionary Trace” method [31] with “Rate4Site” program [32] to construct phylogenetic trees; the use of Bayesian method instead of “Maximum Likelihood” [33] to calculate conservation score; the replacement of Consurf-HSSP [34] database with a more reliable Consurf-DB [35] of pre-calculated amino acid (AA) entropy scores. Such implementations provide extra precision but the methods behind them are sometimes difficult to grasp for an average user. Therefore we report an alternative tool for finding functional important residues in complement with existing tools. The concept behinds our CRC web-tool is simple, it only requires one iteration of PSI-BLAST to build the MSA and calculates the conservation score at each AA position of the MSA. Importantly, CRC allows the user to specify an E-value threshold in making the alignment, an option that was not available in Consurf at the commencement of our analysis. Two protein structures are included in the “Worked Examples” section to demonstrate the use of our tool and may help users to interpret their results.

5.2 Implementation

Methods used in the CRC are from two sources: PSI-BLAST and the C-trident formula. Both were coded in our main program written in the Perl scripting language. The program starts by accepting a PDB ID with a chain identifier as the input, for example: 1a0iA. The ATOM and SEQRES records from chain A of 1a0i are extracted and are written into a temporary file. The
three-letter AA residue code of SEQRES records is converted to its one-letter equivalent. This converted sequence is then used to retrieve close homologues via PSI-BLAST for one iteration (default value) using the NCBI non-redundant database, which includes all non-redundant protein sequences from GenBank CDS translations [36], PDB [5], SwissProt [37], PIR [38] and PRF (Protein Research Foundation) protein sequence database. The choice of one iteration is to minimise sequence retrieval time and to simplify the submission procedure. The number of iterations can be adjusted in our original program.

Two E-values are required in PSI-BLAST and by default both are set to 10e-20. The first run of PSI-BLAST is basically a BLASTp search and any sequences that are below the first E-value threshold are stored. If an iteration number of 2 or higher is given, a PSSM profile would be built from the stored sequences and the profile would be used to fetch more related sequences until no new sequences are detected. Any sequences that are below the second E-value threshold are included in the MSA from PSI-BLAST.

The aligned sequences together with the query sequence are served as the input for the C-trident formula [21]. Three scores (each ranges between 0 and 1) are calculated for each AA position of the MSA and are incorporated into one single conservation score, represented by the formula, C-trident(x) = (1-t(x))α(1-r(x))β(1-g(x))γ, where t is the normalised symbol diversity, r is the normalised stereochemical diversity and g the gap cost. The exponential weights α, β and γ are set to 3, 3, 1 respectively since they performed best in improving the identification of enzyme active site location when sequence profiles were included in our previous enzyme/non-enzyme study [1]. Normalised symbol diversity is a measure of how likely an AA varied in its position, calculated using Shannon’s entropy [39]. The normalised stereochemical diversity uses the BLOSUM-62 substitution matrix to map the diversity of AA properties. The gap is the fraction of the number of gaps in an aligned column. More details of the C-trident formula can be found in the original publication [21]. The file containing conservation scores for each residue of a protein is then matched to the corresponding ATOM records by replacing the B-factor (temperature factor) column with the conservation scores. Thus, a PDB file with residue conservation score between 0 and 100 is generated. (Note: The original conservation score was ranged between 0 and 1 but we have adjusted the magnitude because a few molecular graphical viewers were found to be not interpreting the conservation score correctly.)
5.3 Usage

To submit a job, three pieces of information are required: the user’s own PDB file or input of pre-existing PDB IDs; an E-value for PSI-BLAST search and the user’s email address. Upon submission, all submitted details are validated via a php script. If any information is invalid, for example: an incorrect PDB format, the user will be informed to make a modification (Format details are described in the File Format section). Upon successful submission, the user can either wait for the results by following to the next page or wait for our email when the job is completed. Each job is secured with an unique identifier (ID) and is available to the user and the web administrator only. Calculation for a single protein chain of 350 AA takes approximately 5 minutes with about 36 homologues (including the query sequence) being retrieved. The time may vary according to the protein size and the traffic of the server. The resultant protein structure is displayed with Jmol [40], with the AA coloured in different shades of red and blue representing high and low conservation scores respectively (Figure 5.1) and the white colour indicates the mid-point of the two extremes. A link to the PSI-BLAST log file is available to the user for checking the retrieved homologues.

![Conserved Residue Colouring Webtool](image)

*Figure 5.1: Output of CRC*
5.4 File Format

The user is allowed to choose either to upload their own PDB file with a maximum file size of 2 megabytes or input of up to five valid PDB IDs in a single submission (Figure 5.2). This restriction avoids a heavy load to the server. The submitted structure should follow the described file format showed in figure 5.3. The file name format, for example: 1abcA.pdb, should be a 4-letter PDB ID (begins with a digit between 1 and 9 and three digits/lowercase alphabets), followed by the chain identifier in uppercase with the file extension of .pdb in the end. Only pre-existing PDB IDs are valid, any PDB IDs being made obsolete are invalid. Our PDB files and obsolete list are updated weekly, which follow the PDB website.

![Image of upload interface](image)

*Figure 5.2: A screen shot showing the input interface*
5.5 Worked Examples

5.5.1 Example 1

We have chosen two examples to demonstrate the potential of this web-tool. PDB ID, 1a0i is an ATP-dependent DNA ligase from bacteriophage T7 [41] and is 348 AA in length. From the Catalytic Site Atlas website, the reported functional part is marked by the sidechain of residue LYS-34 (single letter K represents lysine), which is a common residue found in the KXDGXR motif that is reported to be the most conserved area in DNA ligase [42]. E-values of 10e-10, 10e-20, 10e-40 and 10e-80 have been run for the protein. As the E-value becomes smaller, PSI-BLAST picks up fewer sequences, thus the conservation score for each residue is normalised with fewer sequences, which pushes up the conservation scores and more residues are likely to be shown as highly conserved (Figure 5.4). Our result showed LYS-34 has quite a low conservation score even at E-value of 10e-80, as well as other residues within the motif. This is a surprise to us because it is supposed to be one of the most conserved residues. Closer examination reveals that a quarter of the homologues were actually of shorter length (~180 AA) and the segment of KXDGXR motif is absent. This is the main reason why LYS-34 and other residues within the motif gave a lower conservation score. This was not the case in Consurf since it has a scheme to build consensus sequence from the phylogenetic tree before building the MSA, which could eliminate very close homologues. Different weights are also
given to each sequence and consensus sequence according to their evolutionary distances to the query protein. Therefore Consurf is able to make a correct prediction of conservation score at the positions where our approach has failed. The issue could be improved by a scheme to eliminate any homologue with less than 70% in length of the query sequence before going into the MSA. Besides, our program is capable of recognising functional important residues. For the same example, residue ARG-55, GLU-93 and LYS-222 of 1a0i were amongst the top 20 residues (not shown) with high conservation and were close (range from 2.7 to 6.0 Å) to the ATP molecule (Figure 5.5), thus indicating that they may be involved in binding ATP. The original publication [41] confirmed that ARG-39, ARG-55 and GLU-93 form hydrogen bonds with the ribose ring of the ATP. This example suggests our program could be used to identify binding site of small molecules.

![Figure 5.4: Jmol displays residue conservation for 1a0iA using various E-values](image)

(a) 10e-10; (b) 10e-20; (c) 10e-40; and (d) 10e-80. Green circles show in (a) to (d) indicates how the magnitude of residue conservation may increase with decrease in the E-value. Remarks: These figures showed the true image quality of Jmol, since it is popular by its light weighted and browser implementable ability.
Figure 5.5: Display of residue conservation for 1a0iA using an E-value of 10e-20
(a) A normal view. (b) A close-up view to the area with highly conserved residues (ARG-55, LYS-222 and GLU-93), which lie close to the ATP co-factor. Distances between ATP and those named residues vary between 2.7 and 6.0 Å. Remarks: The figures were generated by Molegro Molecular Viewer (MMV) in stead of Jmol because only the iso-surface display is only compatible with the colour setting by temperature factors in Jmol and (2) our program does not include hetero-atoms in the post-processed PDB files.

<table>
<thead>
<tr>
<th>Residue</th>
<th>ARG-55</th>
<th>LYS-222</th>
<th>GLU-93</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>35.141</td>
<td>30.955</td>
<td>35.275</td>
</tr>
</tbody>
</table>
5.5.2 Example 2

The next example, 2zta is a leucine zipper, a motif that is usually found as part of a DNA-binding domain. The submission of chain A using an E-value of 10e-5 to our program revealed that charged residues are more conserved than non-charged residues. Amongst the 15 retrieved homologues, a third were from \textit{S. cerevisiae}; three sequences: 2r32A, 1eboA and 2b9pA (from human, Ebola virus and Parainfluenza virus respectively) were found to have their leucine and valine residues substituted by isoleucine, causing a low conservation at these positions (Figure 5.6). Figure 5.7 shows 2ztaA with the highly conserved Glu and Lys in red and less conserved Leu and Val in blue. This example shows when all zippers were taking into the analysis, charged AAs (D, E, R and K) are more conserved than non-charged AAs, due to swapping between Ile at Leu and Val positions in the human and virus sequences. But when only a particular protein family or organism such as \textit{S. cerevisiae} is considered, both charges and non-polarity are highly conserved.

\textbf{Figure 5.6: Sequence homologues retrieved using 2ztaA via PSI-BLAST with one iteration}
Figure 5.7: Display of leucine zipper, 2ztaA

(a) Conserved charged residues and their sidechains are shown in red and light red. (b) Least conserved non-charged residues and their sidechains are shown in blue.
5.6 Limitation

Our program could fail to give reliable conservation scores if no close homologues are found in the MSA. The current version is limited to single protein chain submission.

5.7 Conclusions

Here we present an online tool, the Conserved Residue Colouring Webtool for searching evolutionary conserved residues in the protein and display their level of conservation on the protein structure. We have shown that the tool is useful in identifying functional important residues in our examples. The algorithms we employed are simple and the submission is made easy for general users. Although there are limitations in our tool, they only affect proteins that do not have any closed homologues in the database. Future improvement should include (1) a better way to display the residue's information when an atom of the structure is selected by the user. Currently, this information is displayed at the bottom left corner of the user's browser. An ideal solution would be an implementation of a floating text box or static text box near the Jmol molecular viewer; (2) A file containing the top 20 most conserved residues may benefit the user from searching in the PDB file; (3) A user-defined iteration number for PSI-BLAST; (4) Hetero-atoms could retain in the post-processed PDB file, so that user may check the conserved residues together with nearby ligands; (5) A sequence selection tool for the user to select homologues to build their custom MSA.
5.8 References


CHAPTER 6 - Conclusions

Our investigations applied computational methods to visit problems including enzyme/non-enzyme prediction (Chapter 4) and soluble proteins/inclusion bodies discrimination (Chapter 3); exploring characteristics that favour the stability of proteins in different subcellular compartments (Chapter 2); and the application of evolutionary conserved residues in proteins (Chapter 5). The aim is to improve the understanding of how protein structure maintains its stability and solubility in the environment and to identify features that would indicate its function.

In an attempt to refine our current method to discriminate enzyme from non-enzyme, we found that our new charge schemes (ISS and PKB), although giving separation, are not as effective as the previous charge scheme (UNI), which is based on the single peak potential given from the largest cleft of the protein structure. However, the results from these new schemes have shown that ionisable groups and their buried pKas can give strong signals, which matched with our knowledge that electrostatic strain from charges are often found in the active sites of proteins. Using single peak potential method, the UNI scheme can predict AS (within 15 Å) correctly in 129 out of 155 enzymes. The predicted number is limited by the location of the AS and the grid step being used in the method. A smaller grid step (1 Å or less) allows the peak potential being sampled at a closer distance to the molecular surface but also picking up large signals from small solvent holes, which affects the results. Therefore, a patch potential method is used, which collects several potentials within a patch before averaging into a single potential and by lowering the grid step, signals from dipoles can now be taken into account, which can have strong effects to the overall potential. Although the performance in UNI dropped, the STD scheme (real charges) can predict AS (within 15 Å) correctly for 123 enzymes. If we have obtained a better performance in STD, we would have begun our studies with the charge template idea. The idea is to characterise electrostatic strain by using potential maps collected from protein clefts. We believe there are charge and dipole patterns lie within the active sites of proteins, that may be able to separate different kinds of enzymes. Therefore, refining our current approach continues to be part of our future goal and we suggested that a second filter could be built to identify active sites within the shallow clefts of proteins.
By calculating the pH of maximum stability ($\text{pH}[\Delta G_{\text{FU}}(\text{min})]$) of protein structures with subcellular annotation, we found that when these properties were averaged over a subcellular location, they were highly correlated with their subcellular pHs. Further investigation revealed that the underlining correlation came from histidines ($\langle\Delta p\text{K}_{\text{a}}[\text{His}]\rangle$, which is the average value of all $\Delta p\text{K}_{\text{a}}[\text{His}]$ from all proteins in one subcellular compartments) and the locations of histidines on the proteins. Within a subcellular location, we have shown that there is a higher proportion of proteins with $\text{pH}[\Delta G_{\text{FU}}(\text{min})]$ on the acidic side if they are from vacuole and lysosome or on the basic side for peroxisomal and mitochondrial proteins. The acidic $\text{pH}[\Delta G_{\text{FU}}(\text{min})]$ in acidic compartment reflects a high and positive pKa of histidine while lower pKa values are found in less acidic or basic compartments. This showed a trend to reduce buffering power associated with histidine in each location but this movement is small.

Since the buffering nature of histidine was observed and we know the pH-dependence of folding energy of proteins around the neutral pH is small (~5kcal/mol), we hypothesised that at any given time in the cytoplasm, a sub-population of proteins would be unfolded, transiently exposing a subset of histidines during the unfolding process. Proton exchanges for this subset of His may slow down or diminished if their pKas have shifted in the absence of the electrostatic interactions from their folded state. Thus, delaying the refolding process, which may lead to poor solubility and may trigger aggregation with nearby proteins of similar condition. This led us to launch our investigation with soluble and insoluble proteins. By comparing their $\Delta p\text{K}_{\text{a}}[\text{His}]$, we hoped to find evidence that the magnitude of $\Delta p\text{K}_{\text{a}}[\text{His}]$ or its derived features may be related to solubility. Although we did find some evidence of histidine ionisation properties correlating with solubility, it was believed that this results from other charge effects that was in turn altering histidine pKas. In particular, charge and non-charge patches are important. We found that the size of the largest non-charged patch can separate soluble-prone and aggregation-prone proteins with a maximum accuracy of 0.74. This finding was in agreement with the general view that aggregation is often caused by hydrophobicity of the proteins. The size of the longest non-charged sequence was also found to be correlated to aggregation-prone proteins but their contribution to aggregation may be small when they were in their folded state. However, partial unfolding might occur in the region of a protein with a high unfolding propensity, and also aggregation if this region is also hydrophobic. Findings from previous work of hyperthermophile proteins suggested that their higher charge percentage and smaller non-
charged patch (compared to mesophile proteins) may avoid protein aggregation (via hydrophobic patch interactions) at high temperature for hyperthermophiles. Following these thoughts, we believed it is possible to increase solubility by breaking the large hydrophobic patch into smaller non-charged patches, which would involve the substitutions of non-polar groups with charged groups. Residues with small sidechains such as Ala or Val would be suitable to be replaced by Glu or Arg, since their longer sidechains can attain higher solvent exposure more readily and have more flexibility. And for protein stability, our results suggested that histidine locations are highly related, indicating an implementation to a correct site on a protein could enhance its stability. However, the current study did not investigate this further, therefore the identification for these histidine locations and their nearby micro-environments will be in the future plan.

We have successfully made our in-house program for displaying amino acid conservation in 3D, available to the public via online access. The program was built for the identification of evolutionary conserved residues that may be of functional importance via the use of PSI-BLAST to retrieve sequence homologues, followed by the calculation of residue conservation. The web-tool allows the user to submit PDB structure and displays residue conservations on the protein surface in the colour shades of red, white and blue, which indicates their magnitude of conservation from high to low. Examples have been given to demonstrate its use and limitation. Currently, the residue conservation value is limited by the homologues selected from the multiple sequence alignment. Our example showed if the retrieved sequences have missing residues, it could affect our results. Thus, we suggested a scheme to eliminate sequences with shorter lengths or an option for the user to remove unsuitable sequences manually in the future version. Also, an implementation of a text box for residue information would benefit the user from searching functional important residue on the 3D structure.

Lastly, our investigations have taken us on a journey across the cell. We recalled from the introduction that a semi-permeable membrane is essential as the cell’s protective layer and act as the first defense for any changes in the external environment. This lipid-bilayer is home for several membrane proteins, which regulate the flow in and out of protons, ions, small molecules and metabolites. The maintenance of pH and ionic strength is key to the functional aspects of the cytoplasm because lots of biochemical reactions happen there. These conditions are also
regulated in the subcellular compartments and are often related to and required for their function. This level of protection and regulation created the platform(s) for proteins and enzymes to adapt and function. It was showed that in general, proteins tend to have their isoelectric point (pI) moved away from their environmental pH, possibly to avoid protein aggregation. There must be a slight excess of positive or negative charged groups on the protein surfaces (sequence derived pI shows similar value) to make them soluble. One experimental study showed that Asp (acidic, negative charged) and Ser (polar) can enhance the solubility of proteins. Our solubility study has showed that the surface of a soluble protein tend to have the largest non-charged patch smaller than the one from aggregation-prone protein in general. Other study found that higher content of aromatic residues (Phe, Tyr, Trp) reduces protein solubility. However, Trp is found to have other functions as well, for example, it is observed to be responsible for sugar binding in carbohydrate binding site, which has a stabilising effect on protein structure upon sugar binding. We also showed the accumulated effect of His on the pH[ΔG_FU(min)] of proteins in a subcellular compartment. These above examples indicate that the physico-chemical properties of the amino acid sidechains and their locations are important for the adaptation of proteins in the cell. And for a protein to be functional, it was known that having a large cleft is a strong indication but not a compulsory feature. Our study showed that using the ionisable group derived charged schemes (ISS and PKB) can identify enzymes that were previously predicted as non-enzymes in UNI scheme. This further confirmed that charged residues and their electrostatic strain are useful for discriminating enzymes/non-enzymes. Because electrostatic strain is often seen in protein functional sites, other proteins with the same or similar function are likely to conserve the electrostatic strain pattern or those crucial residues. This is the purpose of our web-tool and it helps identifying functional importance residues for unknown proteins. And hopefully in the near future, it could recognise strain pattern when our AS prediction method is finally improved.
ADDITIONAL FILES (Available on the web)

Additional File 2-1: Dataset used (Chapter 2)
Description: Excel file containing the PDB IDs, their chain identifiers, and subcellular annotations, of proteins used in this work. The data follow the description given in figure 2.1.
http://dl.dropbox.com/u/21860511/PhD_Thesis_Additional_Files/Additional_File_2-1.xls

Additional File 2-2: Separate Dataset used in the pH Stability Test (Chapter 2)
Description: Excel file containing information for proteins used in a test of pH stability predictions.
http://dl.dropbox.com/u/21860511/PhD_Thesis_Additional_Files/Additional_File_2-2.xls

Additional File 2-3: Program for pKa Calculation (Chapter 2)
Description: Fortran program for protein pKa calculation and analysis.
http://dl.dropbox.com/u/21860511/PhD_Thesis_Additional_Files/Additional_File_2-3.zip

Additional File 3-1: Datasets used (Chapter 3)
Description: Excel file containing original tables that our datasets derived from. Also containing the PDB IDs for our final datasets.
http://dl.dropbox.com/u/21860511/PhD_Thesis_Additional_Files/Additional_File_3-1.xls

Additional File 3-2: Raw Result Data for Datasets N1 and N2 (Chapter 3)
Description: Excel file containing raw result data for N1 (111 SPs) and N2 sets (57 IBs).
http://dl.dropbox.com/u/21860511/PhD_Thesis_Additional_Files/Additional_File_3-2.xls

Additional File 4-1: Datasets used in the First Study (Chapter 4)
Description: Excel file containing datasets and raw result data for 54 CB enzymes, 37 CB non-enzymes and 36 PBPs used in the first study of this work.
http://dl.dropbox.com/u/21860511/PhD_Thesis_Additional_Files/Additional_File_4-1.xls
Additional File 4-2: Enzyme Set used in the Second Study (Chapter 4)
Description: Excel file containing dataset and raw result data for the 155 enzymes used in the second study of this work.
http://dl.dropbox.com/u/21860511/PhD_Thesis_Additional_Files/Additional_File_4-2.xls